Business English Trainers

Series Editor: Eric Baber

# Testing and Assessment in Business English

Judith Mader

Professional

CD-ROM

Cornelsen

Series Editor: Eric Baber

# Testing and Assessment in Business English

Judith Mader

**Professional**
Business English Trainers

Cornelsen

**Verfasserin** Judith Mader

**Serienherausgeber** Eric Baber

**Redaktion** Stephanie Hempel

**Projektleitung** Murdo MacPhail

**Projektkoordination** Anna Batrla

**Bildredaktion** Uta Hübner

**Produktionsmanagement** Kathrin Eitdorf, Johanna Strodt

**Layout und technische Umsetzung** zweiband.media, Berlin

**Umschlaggestaltung** hawemannundmosch, Konzeption und Gestaltung, Berlin

**www.cornelsen.de**



1. Auflage, 1. Druck 2011

Alle Drucke dieser Auflage sind inhaltlich unverändert
und können im Unterricht nebeneinander verwendet werden.



Druck: CS-Druck CornelsenStürtz, Berlin

ISBN 978-3-06-520080-6

Inhalt gedruckt auf säurefreiem Papier aus nachhaltiger Forstwirtschaft.

# Foreword

## Testing and Assessment in Business English

Perhaps more than ever, assessment is now an integral part of the learning process to the extent that it often features at the beginning of, the end of and at multiple points within a learning path. Whether you are using a placement test bought off-the-shelf or created in-house; a weekly test that you have set yourself; an exam published by one of the large exam houses; or an end-of-term institutional examination, assessment is part of every Business English trainer's and language learner's life.

As a result, it is important to both get an assessment right and prepare your students for it appropriately. It may sound banal, but when creating an assessment yourself it is vital that you test what you mean to test (and you might be surprised at how many home-grown tests get it wrong in myriad ways). Similarly, teaching learners both the language and the skills they need for an exam they may need to sit is one of your key responsibilities as a trainer.

On the other hand, it is important to strike a good balance between teaching English and teaching exam skills. Too heavy an emphasis on assessment itself to the detriment of language skills runs the risk of demotivating or boring your learners, and not preparing them adequately with the language skills they need in order to succeed in their chosen career or life path.

This book aims to be a helpful guide for Business English trainers in navigating all of these issues. It provides information on top-level elements of what makes a good assessment and some of the pitfalls to avoid. It then drills down into specific areas of assessment including discrete question types and their relative strengths and weaknesses, and addresses the testing of different language skills and areas. For Business English trainers in Europe, there is also an invaluable discussion on the CEFR, the Common European Framework of Reference for Languages, and how it impacts on assessment and teaching. The overview of published exams should also help you give your clients advice on which of the published exams are the right ones for their specific needs.

The CD-ROM contains checklists to help you apply what you have just learned in the chapters, ensuring you have applied the main criteria relevant to the type of assessment you are constructing. Also on the CD are adaptable worksheets and guidelines to facilitate an in-house training workshop for yourself and your colleagues, to ensure everyone within your institution is up to speed on all things assessment-related.

We hope you enjoy the book and find the materials interesting and stimulating.
Happy reading!

**Eric Baber**

Eric Baber has a wide range of experience in the area of Business English teaching including as a teacher trainer, author, and tutor on online language and training courses. He is currently President of IATEFL, the International Association of Teachers of English as a Foreign Language, and is a regular speaker at ELT conferences around the world.

# Table of contents

*"… in order to be useful, any given language test must be developed with a specific purpose, a particular group of test takers and a specific language use domain (i. e. situation or context in which the test taker will be using the language outside of the test itself) in mind."*

Bachman, Lyle/Palmer, Adrian (1996): Language Testing in Practice. OUP, p. 18

# Introduction

This book is intended for trainers of Business English (BE) who want or perhaps have to deal with issues connected with testing, assessment and evaluation. They may be teaching in a range of situations such as companies or universities, universities of applied sciences or further education colleges. What they will have in common is that they are primarily concerned with teaching English for job, work, occupational, business or professional purposes. This means that they are working with adults or young adults who are either already in work or are preparing for it. The learners' interests (even if they do not make this explicit) are in performance in English, use of English and successful communication. Although most of what is in this book applies to all language levels, the focus is not on beginners, false beginners or low levels, but rather on what used to be commonly known as *upper-intermediate* or *advanced* and now called B2. These are the general parameters of this book, although much can be applied to other situations where language testing is an issue.

If you are reading this book, you have probably been asked either to produce a test or some form of assessment or to select from existing available tests. When BE trainers get involved with language testing, it is generally in order to do one of these things. The tests they have to produce can range from placement tests to classroom tests, end-of course tests or exams (in the case of a university). The tests they have to select may be published placement tests for use with course books, Internet-based tests or external exams run by international examination boards.

**What is it?**

The term *testing* is usually used to refer to fairly formal specific situations, whereas *evaluation* is used more for a global judgement of language skills, but also of a course or programme and *assessment*. *Assessment* often means a less formal judgement, made through observation, of size and value. In the context of teaching, this means something taking place over a period of time with the intention of improving learning.

## Focus

As the most common way of assessing learners' ability in a language has traditionally been using tests. This will be the main focus of the book, although assessment and

evaluation will be mentioned as well. A test is usually seen as something which takes place at a specific time with a specific group of learners for a specific purpose. It is governed by a set of rules and produces results which are used for some purpose, either to benefit the learners in some way or to provide information which is used as the basis for making decisions.

This book is primarily about language testing for BE trainers, and I hope to be able to make the topic as easy to understand as is possible and to deal with the practical issues facing BE trainers concerning testing in their everyday work. Any BE trainer who has started to become involved with testing seriously will know that it is an extremely large field with very many academic and specialized books available. There are several groups and mailing lists as well as journals dealing only with language testing. This is in addition to the treatment of language testing as a topic in general ELT and BE publications.

**To think about**

This book is not intended to offer a great deal of theoretical background to testing and assessment. Instead, it focuses on the realities of implementing exams and tests.

Testing is not theoretical, although it may sometimes seem that way. The theory on its own is of little value to BE trainers without the practical application which follows it. Much of the discussion and controversy on testing issues is between theoreticians and practical test setters. Their approaches differ, as they see the subject of testing from two different and thus opposite (but not opposing) sides of the field. While it is certainly extremely valuable to know something about the theory behind testing, I am assuming that the average BE trainer and reader of this book will have neither the time nor the inclination to go into the theory of testing in great detail. Most probably you will want to produce a good test when you have to, or maybe even (only) to select the best test for your learners from the range of tests available on the market.

**This book aims to condense the most important, and in particular the practical, aspects of testing and make them accessible to practising BE trainers so that they can do the best job possible when producing, evaluating or selecting a test.**

There are aspects of testing which are not covered in any depth in this book and some which are not mentioned at all. For those who want to know more, a reading list is provided with suggestions for further reading. All of the books on the list have references to other books. The difficulty is sorting out what you really need to know from the literature available. This book is intended to provide some help with this.

## Terms

In this book, I use the term *trainer* to refer specifically to BE trainers and *teacher* if I mean the term in a more general sense. So far, in this introduction, I have used various terms for the production of tests (*test setting, test construction, test writing* and *test production*). In the following chapters, however, I use only *test setting* and *test setter* as the terms for this activity and person. When learners are doing a test or it is this role which is important in the context, they are referred to as *test-takers*. Otherwise they are referred to as *learners* or (when a university context is specifically referred to) as *students*. The term *examiner* is used for the person conducting and marking an oral test and the term *rater* for the person marking a written test. This is in particular when judgement is involved and not merely deciding whether an answer is right or wrong. Other terms with a specific meaning are explained in the relevant chapters or in the glossary.

## Teachers and testers

On the subject of teachers and testers, it is not necessarily the case that the best teachers are the best producers of tests. However it is often assumed by non-language experts, such as personnel managers or heads of departments, that if they are good at language

teaching then they can produce language tests just as well. But while these trainers may know a great deal about teaching BE, they rarely know as much about testing and sometimes know very little. Just as general medical practice and surgery are really two different branches of one subject, with different professional standards, so are teaching and testing.

Being a good tester does not involve the same qualities as being a good teacher. For instance, it could be argued that there is less need for creativity, charisma and other communicative skills and a greater need for a systematic and a possibly less creative approach. Whatever the reasons may be, teaching and testing are not the same thing, and there is no reason why one person should like or be good at both. Thinking about what you really enjoy about teaching may help you to decide how much you want to take on the job of testing in a professional way. Your teaching experience will certainly help you when designing and setting tests but you will also need qualities such as being able to work systematically and thoroughly to complete the job of setting a good test.

If you really feel that setting tests is not one of your strengths at all, and you have the chance to pass the work on to a colleague who finds it enjoyable, this may make more sense. However, it is important, just as with diagnostic medicine and surgery, for teachers and testers to know something about what the other group of professionals does. It may be for this reason that you are reading this book. You may in fact find out, when you know something about it, that testing appeals to you and you want to try out what you have learnt.

## Some important points

There are several basic things which language teachers should remember about language testing. All these apply just as much to BE.

**The most important thing to be said is that producing (constructing, writing, setting) a good test is not simple.**

This book should make it easier for you to understand what is involved in setting a good test, but it will not do the work for you. You may feel that you have been setting tests for years and they have served a purpose, but possibly not entirely satisfactorily. This is because it is very easy to produce something that looks like a test, but which in fact is nothing of the kind. This is the first thing that readers of this book should be aware of. Something which only looks like a test may impress people who know little about testing, but it will not do what you want it to. A test which is really good, i. e. does its job properly (and we will go into this in detail), takes time, effort and resources (people and, to a lesser extent, money) to produce. This should not discourage you but rather help you to approach the matter in a sensible and practical way.

**There is no such thing as a perfect test.**

This may seem to contradict what I say above, but although it is not easy to set a good test, it is unrealistic to expect perfection. When people and language, both of which change with the circumstances and often behave unpredictably, are involved, nothing is easy. The most a test setter can do is to try and come as close to a perfect test as possible. While you may never set a perfect test, you can deal with the shortcomings of tests (your own or published tests) as realistically as possible so that both you and your learners get the most out of the testing experience.

One important thing is not to be discouraged if your early attempts at producing a good test are not successful. Even if there is no such thing as the perfect test, there are certainly

good and bad tests. Test-setting, like teaching, improves with practice and many aspects become routine. The more routine there is in the production of a test, the easier the process and the better the test will become as the more you can concentrate on the difficult parts.

**A test must do some good.**

This may seem obvious, but it may be worth thinking about whether this is always the case. Consider your answers to the following questions about the tests which you have used or experienced up to now.

- Do learners understand why they are doing the test?
- Do they feel very nervous before it?
- Does it negatively affect their view of BE or their attitude to English in general?

If your answer to any of these questions is yes, then your test is not doing its job well. While a test need not necessarily be fun, it should always in some way be a beneficial experience for test-takers.

**Remember**

A test does not take place in a vacuum. It is set for a reason, with a group of (potential) test-takers and takes place in a particular setting or context.

This will involve the target group, your own learners perhaps, being involved in some way in the setting of the test, even if only in your mind. If you know what you are doing, then you can convey this to test-takers. The more the test-takers understand about the test, the better they are likely to perform in it.

**Setting a test takes time.**

This means that you should always allocate a reasonable amount of time for the job. Don't expect an end-of-term exam which you have hurriedly typed up the evening before your students have to sit it to produce results you are happy with and which are useful for the students. Even very experienced test setters cannot set tests in this way. You may have saved time setting the test, but you are almost bound to have overlooked one or more things. There may be several possible answers to the questions, you may not have thought of all the answers test-takers can give, there may be typing or spelling mistakes (which should definitely be avoided in a language test), the test will most likely prove difficult to mark and the time you have saved setting it will have to be spent marking it and working out a way to deal with the mistakes you have made. In fact, it may be because you have experienced something precisely like this that you are reading this book.

The whole process of testing (setting and marking and all the other things involved) takes time. It is only how this time is divided up which varies – according to the sort of test and what is being tested. For example, multiple choice questions are easy to mark, but it is very difficult to set good ones. Essay questions are relatively quick and easy to set, but require instructions and criteria in order to mark them. These need to be developed and may take training and so be time-consuming to apply. There are so many things to consider that setting a test is not something that can be done at the last minute.

Start thinking about the test you want when you plan the course, and bear the test in mind while the course is going on. Make it clear to anyone who asks you to set a test that you will need time for this if the test is to be any good. Make sure that everyone involved knows that it is a worthwhile investment of time, manpower and money.

This relates to the next point.

**Never set a test on your own.**

Always use colleagues, native speakers and learners to help you decide whether your test is a good one. They will of course see the test from different perspectives, and some will

be better at, for instance, proof-reading, others at formulation of instructions and texts and others at answering the questions, to name but a few aspects. Take all their comments seriously and if possible discuss them before you decide which comments and suggestions you should adopt and what changes you should make.

If you do not want to impose too much on other people, allocate tasks and think about how much these can be a part of teaching. Parts of the test can be tried out by colleagues in their lessons (assuming their learners are not going to be taking the test). They will probably be grateful for "free" no-preparation material and willingly give you their and their learners' feedback on it. You can reciprocate by giving the learners feedback on their results.

**Remember that your tests (or rather their results) may affect your learners' lives and careers.** Test constructors are not surgeons and it is unusual (but not out of the question) that a test-taker's survival depends on the results of a test. However it is easy to forget that the results of a language test may be extremely important for the individual involved. The results of an end-of-term exam may affect the final grade in a degree course, the results of an end-of-course test the chance of a higher salary or a more interesting job and the results of a language test for citizenship a work permit or stay in a foreign country. Even something as seemingly simple as a placement test, if badly set, may lead at best to learners being put in the wrong course and having to be re-placed at some inconvenience, in the worst case to their having to spend several lessons or longer in the wrong course, leading to frustration all round. Even a bad result in a class test may affect learners' self-esteem and their motivation to go on learning which in turn may reflect on the teacher. If you are not actually doing a test for a reason which is important (even in a small context), it may be worth thinking about why you are doing it at all. Just thinking about the reason for setting or giving a test will help to make your test better.

Although these points may seem intimidating, working professionally with language tests is made easier by adopting a systematic approach to the subject. There are many similarities to designing a course or choosing a course book. Both these areas, which should be familiar to BE trainers, have to be approached systematically, and testing is no different. The basic principles described in this book will help.

## This book and how to use it

The book consists of this introduction and 16 chapters, a list of suggestions for further reading with relevant websites, a glossary, and a CD-ROM.

Each chapter starts with a quotation and some introductory questions to give you an idea of the content of the chapter. There are boxes in the margin providing cross-references to other relevant chapters as well as boxes which are headed **Remember, Did you know? To think about** and **What is it?**. There is an example of each of these in this introduction. Each chapter ends with further questions for you to consider and answer in relation to your own testing context.

**Chapter 1** is essential reading as it will help you to understand what a good BE test is and what you need to decide before setting or choosing one.
**Chapter 2** is about different types of tests and how they differ.
**Chapter 3** is about question and task types and what you need to consider when setting them.
**Chapter 4** is on specifications which are produced and used by all good test setters. Specifications are used to answer questions about the test which may be asked by those involved.

As the question of language level is always an important one, **Chapter 5** has been devoted to the Common European Framework of Reference for Foreign Languages (CEFR) and its increasing importance.
As many BE trainers may have to choose tests for their learners rather than set them themselves, **Chapter 6** is about the main BE tests available on the market and how you can select the best one for your learners.
Whether and how much you read of **Chapters 7 to 12** at any given time will depend on what you are testing, as each of these chapters goes into one of the following specific areas: reading, listening, writing, speaking, vocabulary and grammar. You can select these chapters according to your particular interests and needs.
The final chapters again apply to different sorts of tests and will also be useful when selecting a test. Each chapter has a different focus.
**Chapter 13** is about the designing of criteria and the setting of scores as well as using these for marking and grading.
**Chapter 14** is about various practical issues, including the stages of test setting and test administration
General issues such as fairness and washback are dealt with in **Chapter 15** along with other forms of assessment.
Finally, **Chapter 16** gives you some ideas on how best to prepare learners for exams and tests.

It is not always possible to avoid using technical terms in the text, so those which may be unfamiliar are explained briefly in the **Glossary**.

Clearly a book like this can only cover a certain part of the vast field of testing, so there are some suggestions for **Further information**.

The **CD-ROM** contains checklists relevant to the topics of the chapters. There is generally one checklist for each chapter. They can and should be used every time you carry out the activity the checklist refers to (e. g. setting a test, using the CEFR, writing test instructions). The CD-ROM also includes some templates, for example for setting tests and creating marking schemes. These can be adapted to suit your decisions on marking and scoring. There are also some ideas and material for running a short workshop with colleagues. Although you will not be an expert after reading this book, and a workshop with colleagues will not replace attending a course on testing or inviting an expert to advise you on testing issues, the workshop approach is a very useful one when you are setting tests, especially if you do not have a great deal of test setting experience.

**Did you know?**

The CEFR has been adopted officially and unofficially not only all over Europe, but in many countries outside Europe too.

## One final point

Testing, like teaching is something to be taken seriously. This book should help you do just that without exposing you to too much theory or making the whole issue too difficult. You may even find that you too join the strange breed of test-lovers to which the author of the book and many others belong!

## Where do I go from here?

■ As Chapter 1 is a general overview, it is best to read this first. You can then proceed to the chapter which applies most to your situation or interests you most. You may have to choose a test, set a particular type of test or want to test a particular skill or aspect of language. If you are not selecting a test from the range of publicly-available tests, but intend to set your own, it will probably be best to think about test types, question types and specifications, which are the subjects of the following chapters. ■

*"Language tests are not devised to find out the degree to which learners are deaf or blind, whether they know the name of the Prime Minister or whether they believe in God!"*

Pit Corder, S. (1973): Introducing Applied Linguistics. Penguin, p. 360

# 1 What is a good Business English test?

## On good tests, validity, and dispelling myths about testing

- Why test at all?
- What theory is necessary?
- What is Business English?

Let us assume that you have this book in your hand because you are faced with the task of setting a test or of choosing the best test for a particular group of learners. This may be a part of a training strategy for your institution which you have been asked to devise and implement but it is only the testing part which will be addressed here.

Let us also assume that you are a BE trainer (we will come to what we mean by BE later) who is not an expert in testing theory and is really more interested in teaching than testing. It may well be the case that testing is best left to the experts, but it is often the chalk face teachers who are entrusted with the task of setting and choosing tests and who carry it out with varying degrees of success. As you are reading this book, we can assume that you want to do the job well and are prepared to invest a certain amount of time and energy in it, but that you would probably want to minimize this.

### 1 Questions

For the purposes of this initial chapter, it is not important whether we talk about test setting or test selection, as the questions to be asked before doing either are much the same. This book will ask (and go some way towards answering) the most important

questions. Test setting rather than test selection will be the main focus, as setting a test is definitely the more complex task and is not to be taken too lightly. As far as selecting a test is concerned, if you get close to mastering the art of test setting, you will be ideally equipped to look critically at existing tests and to choose the best one for your purposes.

This book has been written for BE trainers but only addresses them in their function as test setters or selectors. In many cases, it is easier to place yourself at one remove from the test-takers and imagine that another teacher has been, is, or will be their teacher. This will give you the feeling that others will benefit from your decisions and you will not be able to cover up or compensate personally for any negative effects on test-takers. If the test which you have set is to be used by others and is not up to much, this will reflect badly on you. Imagining that this is the case will give you at least an extrinsic, in addition to an intrinsic interest (we hope!), in doing the job properly.

**Did you know?**

There is a school of thought which says you should not set a test for a group you know!

Although it is important to know the target group for the test, it is not necessary to know the individuals personally. You may know them (especially if you are their trainer) but what is important to you is what they have in common, in terms of their needs and as far as the test is concerned.

Before going on to consider particular types of tests and indeed particular tests and exams available on the market, it is necessary to ask a fairly large number of questions. This may at first glance seem tedious, but only by thinking about the answers to these questions, can you really save a lot of time and effort when actually setting the test. The biggest mistake most teachers make is to launch into test setting with insufficient preparation. This can lead to a lot of wasted time and effort. It makes sense to invest time in thinking about the test (wherever you do this – in the shower, on the train) as this will lead to a far better test when you come to actually set it.

The answers to the questions posed here will lead you to the answer to the main question in the title of this chapter. After that, you can decide which aspect of testing you want to deal with and continue from there.

The task of test setting is best approached by first asking two questions. The questions apply to any test.

1 Why am I setting this test?
2 What do I want to test?

## 2 Why am I setting this test?

Which, if any, of these answers to the first question apply to you?

- My learners want to do a test in the next lesson.
- A test is a good way of keeping the class occupied.
- The rules of my institution call for a test/exam.
- I need to find out what my learners already know or can do in English before starting the course.
- I want to know what I have to repeat or teach in the next course.
- I want to see how my learners compare with others in the company/institution.
- ...

Your own answer may vary slightly but will probably fall into one of these broad categories. Basically the first two reasons are not particularly good ones for setting tests. If one of these is your main reason for setting the test, then you can do more or less

anything you want, as it will pass for a test even if it is by strict standards not one. Also it will be only for your own consumption (and that of your learners, of course) and will probably do little harm (or good).

If you are setting a test for the other reasons then you will have to approach the matter more seriously and ask yourself one or two more questions before proceeding.

The answers will take you into the process of test construction. This means venturing a little way into testing theory (a subject not actually as daunting as many believe), and coming to grips with validity, reliability, practicability, impact and objectivity. Of these validity is by far the most important.

## 3 What is validity?

Validity is basically the degree to which the test is testing what it sets out to test. For example, an English test should test English, a Business English test Business English, a test of reading, reading and not listening, a test at a certain level, the things applying to that level and so on.

**Remember**

Validity is essential because it is used for making decisions about the test. Validity is the appropriateness and meaningfulness of a test for a specific context, i.e. the framework for what is being tested.

This means of course defining these first, which is not always easy. However it is usually fairly easy to define what each is not.

**Example**

*There were a hundred and eighty delegates at the conference and about 45 representatives from multinational companies. Of the rest 30 were from SMEs, 75 from …*
*Is the following sentence true or false?*
A quarter of the people attending the conference came from multinational companies.
→ Does this test maths or English?

There are many different forms of validity, which relate to BE in different ways. The most important however are, in order of importance,

- Construct
- Content
- Criterion-related
- Scoring

and something will be said about each here and in the rest of the book.

**What is it?**

Validity is nothing new. In 1927 Kelly said "The problem of validity is that of whether a test really measures what it purports to measure" and in 1961 Robert Lado posed the vital question "Does a test measure what it is supposed to measure? If it does, it is valid."

**Construct validity** is related to the view of language your test is based on. This relates in this case to what you consider BE to be. To take a general example, if your view (the construct within which you operate) is that communication is the most important aspect of language, then a test which tests grammatical accuracy will have little construct validity. This is why it is important to think carefully about what you consider BE to be.

**Content validity** is closely related to construct validity but in many ways simpler to understand. To have high content validity, a test must test what it says it is going to test, i.e. a test of English should not include questions which can only be answered with mathematical skills (not as uncommon as you may think) or only be answered by using knowledge of subjects other than English (such as general or cultural knowledge). In terms of BE the content of a test (or a course) may include business knowledge as well as language skills.

**Criterion-related validity** concerns the criteria against which a performance is measured. How well does a test-taker have to perform in order to pass a test or to get a high grade

and how badly in order to fail? In order for a test to be valid in terms of its criteria these cut-off points must be defined clearly. This can only be done by considering the level of the test-takers and the test itself and developing criteria to define these.

**Scoring validity** is related to this, in that the decisions the setter makes in terms of the scores for the test and its component parts must also relate to each other appropriately, i.e. the scores of individual sections must relate to the importance attached to these aspects of the test as a whole. For example, a test of communicative competence in BE should not award the majority of its points for translation or questions of grammatical accuracy.

▶ Chapter 13

Similarly, a test of presentation skills should probably not award the highest number of points for slide layout (unless this is clearly defined as important). Scores and their weighting must be decided on before starting to construct the test.

A good test should also have a theoretical foundation as its basis, e.g. of what the construct, content, criteria and scores mean in terms of the candidate's overall performance in the test and outside. As a test can only deal with one fairly small segment of overall performance, the choice of tasks, texts and questions must take theory into account. This may have been done when making the decision on the construct. A test setter may wish to rely on theoretical background work already carried out by others, for instance on the nature of reading or listening and the skills and sub-skills involved. All good exam boards will have taken the theoretical background into account and provide references to this somewhere.

▶ Chapter 6

Let us return to the questions you will now have to ask in order to make sure your test is as valid as possible.

## 4 What do I want to test?

Your answer to this is presumably "Business English", so we should first attempt to define what BE is before going on to consider what a good test of BE is. There is no standard definition of BE although several authors have provided their definition, and these may be worth consulting.

There are many useful versions of BE, including those given by Nick Brieger, Sylvie Donna and Evan Frendo.

The IATEFL Business English Special Interest Group (BESIG) has discussed the possible differences between BE and English for Business Purposes which apply to level and skills, curriculum and materials development, teaching methods, classroom management and testing. It is not really important what you call it, but what you mean by BE can make your underlying ideas of the English which is actually needed for work today clearer and answer the question "What (Business) English should I be teaching and testing?"

**What is it?**

A *high-stakes test* is one on which a lot depends, for example a university entrance examination, a test of language for awarding citizenship or a school-leaving exam.

Depending on how high-stakes your test is to be, you may want to define BE more formally and consult others on the matter. If however you are setting a small classroom test for one group of BE students, your own definition may be adequate.

### What do you mean by Business English?

Which of the following most closely resembles what you teach and what you intend to test?

**Definition A:** Business English is the English used by people at work and for business in the widest sense of the term. It pre-supposes a basic knowledge of English grammar and vocabulary to which specific business language (mostly vocabulary) is added. This language depends on the learners' areas of work.

**Definition B:** Business English is used to define the language skills necessary for work. It is not so much the grammar and vocabulary but more the skills required for tasks like negotiating, listening to a lecture, reading product information or giving a product presentation. It presupposes a sound knowledge of English (grammar and vocabulary) and concentrates on the appropriate use of this at work and in business. The skills involved will depend on the learners' areas of work.

**Definition C:** Business English is the English used in particular situations, such as ordering, complaining, booking a room or giving instructions. No particular knowledge is required and it can be taught from beginners' level, if the focus is on language (grammar and vocabulary) used at work and in business. The situations will depend on the learners' areas of work.

**Or define BE yourself ...**

- Which of these three definitions is closest to the type of Business English you teach?
- What implications does your definition have on the content of your courses?
- What do the three definitions have in common?
- What will be the implications of the definition you choose for a BE test?

These basic considerations have serious implications for the validity of the test and so are the first things you have to decide. From here you can proceed to a more detailed description of what you are testing, usually known as the test specifications. These will include information on the type of test, intended test-takers, benefits and effects etc.

**Remember**

All good tests need specifications.

▶ Chapter 4

## 5 Features of good tests

### Reliability

Reliability is the correctness, consistency and equivalence of a test. All these must not only exist for a test to be considered useful, they must also be demonstrated in some way. A test must produce reliable results, i. e. the users of the test must be able to depend on them. This applies just as much to a single test as to a series of tests. In the case of a series of tests, the results produced should be comparable.

An international examining board must make sure that equivalent results are produced for all candidates, for example whether they take an exam in spring or in autumn. Similarly end-of-term exams must not favour one group of students over another by being easier or more difficult from one term to the next. It is common for tests in a series to become more difficult as time goes by. Test setters become easily bored by their texts and tasks and in their attempts to introduce more variety into their tests, unintentionally raise the level of the test.

This is one of the main reasons for always having one or more colleagues read, comment on, modify, revise and edit tests, even on a small scale.

**Remember**

The most important thing is never to set a test alone!

**Did you know?**

"The end of the line" means that a test is set and taken and then everything starts again, i.e. no one learns anything from the process.

There may be situations where serial reliability is not such a serious issue, for example if a test is only going to be set and taken once and never again. However the results need to be reliable! It is rare that a test is really "the end of the line" as most tests have some consequences which then have a bearing on their reliability. So even in an apparently one-off test situation, reliability may not be such a non-issue as some may think.

The reliability of a test is measured against the results it produces (even for one-off tests) and against other tests. These may be other tests in the same series or other tests claiming to test the same thing at the same level.

## Authenticity

**Remember**

Test tasks should always be as authentic as possible and at the same time as practicable as necessary.

Authenticity is how the content of the test is related to the use of what is being tested and how this is taught. To take an example unconnected with language testing: a driving test would not be considered authentic if the test-taker were not required to go onto the roads but merely asked to answer questions about driving without having to interact with other drivers in traffic.

What an authentic task in a language test is will depend on what is being tested, in what sort of test and what candidates have been asked to do in class. The authenticity of tests of BE will depend on the definition of BE chosen and the context the test is used in. This also relates closely to "teaching to a test".

Unfortunately many authentic tasks are not practicable for testing purposes (telephoning with real telephones may be one) and can only be simulated in as realistic a way as possible.

## Interactiveness

This is the degree to which the test-taker can interact with the test tasks appropriately so that the test allows performance at the highest level of ability. This means that the test-takers' ability, skills and knowledge should be invested in the test and not in the task.

To again take the driving test as an example, what is asked of the test-taker corresponds almost exactly to what he/she has to do as a driver on the road. In many language tests candidates are asked to perform tasks which they never have to perform outside tests, such as filling in gaps or answering multiple-choice questions. Usually these tasks are chosen for reasons of validity, reliability, objectivity and practicability and, for these reasons, can be considered acceptable as long as the candidates are not over-exerted mentally. The tasks mentioned are simple to understand, are used in many sorts of tests and are often used in course books.

How far candidates can interact with the tasks in the test will depend on their strategic competence, topical knowledge and other things which will have to be considered when setting the test. Test-takers who have been out of the learning process for some time may have difficulty with multiple-choice questions and may need explanation and/or examples. Some types of question or task are not always suitable for teaching but useful for tests. Familiarising learners with these and giving them explanations or examples may be enough to raise the degree of interactiveness of the test to an acceptable level.

▶ Chapter 16

## Practicability

Although many test theorists would prefer not to have to discuss the issue of practicability or feasibility at all, there can be no doubt that it is one of the most important considerations for test setters in the field. Unfortunately it is often the starting point for a test setter. A company needs a test for 20 candidates; it should not take more than 30 minutes and

should only cost €200 to produce. What is being tested and how this is to be done are unfortunately often secondary considerations. The art of setting a test is combining both the practical and the theoretical to make the best possible test.

Practicability depends largely on the testing situations, but in general, the most practicable tests are probably low on validity and reliability. Using video in an oral examination may not be practicable, observing language users over a period of time longer than about 60 minutes is usually not practicable for large numbers of learners, using multiple scorers for written pieces of work is a big step towards greater objectivity and reliability, but may not be practicable (usually for financial reasons). It is worth considering which practical issues are essential or important in the setting of a test and which can be approached flexibly.

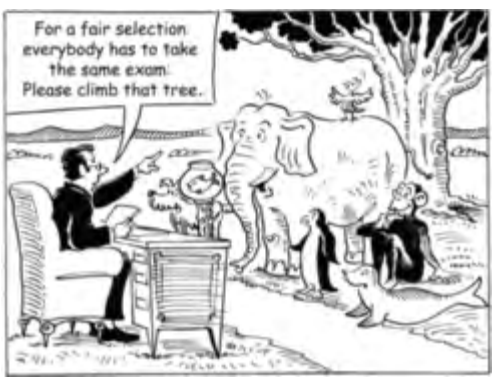
## Impact

The importance of impact as a consideration in test setting will vary considerably depending on the type of test being set. The impact of a test may have little or nothing to do with the quality of the test itself. For example, a language test which is the part of the procedure for obtaining citizenship of a country has an extremely high impact.

You will have to decide how much impact the test will have, how much this will affect you and what sort of test you set. If, for example, you are setting a placement test for a company, it may be adopted by the company, used in all its branches and become the standard test for all language courses for many years to come. On the other hand, it may be used once, then only to be thrown away. The impact may have nothing or everything to do with the validity and reliability of the test. A test with a prestigious name may have great impact but few other qualities. An excellent test may have no impact at all. If your test is designed to show that those candidates who pass it can translate business correspondence, then it will not help the reputation of the test if their bosses realise quickly that holders of the qualification are not able to do this. The impact will be low.

## Objectivity

Lack of objectivity is probably one of the most common criticisms aimed at tests, i.e. that they favour some test-takers over others and that examiners pass and fail test-takers according to their whims. If a test claims to be objective then it should demonstrate this. There are various ways of doing this.




If test questions have specific answers, the setter must be certain that only these answers are really correct. This is done by using expert judgement and checking and re-checking the key.

Open questions must have criteria against which they are marked, scored and graded. Open questions are those which have no one correct answer or set of correct answers. Test setters have to decide which answers to questions and solutions to tasks they accept and to what extent. This involves designing criteria. These should be as detailed as possible and aim to cover as many eventualities as possible.

▶ Chapter 13

Sample answers (both low and high scoring) and reasons for the scores will increase the objectivity of the test. Similarly transparency in all other aspects of the test will also increase its objectivity. This can include format, scoring, criteria, sample questions, practice material, examiner training as well as benchmarking and appeals procedures. This is why specifications are so important.

▶ Chapter 4

### What is the most important thing to consider?

**Remember**
- There is no perfect test.
- No one test can do everything.

All of the above considerations interact with each other. This affects the degree to which they can be applied. For example, a highly practicable test may be low on validity or a completely objective test low on authenticity. It is really up to you as a test setter to decide how much of each you want. This may also depend on what you are testing.

However once you have decided what you want your test to do, you can try your best to make sure that it does just this. It is by no means impossible. First let us dispel some myths about testing in general and language tests in particular.

## 6 Some testing misapprehensions and myths

Misapprehensions about tests abound. Some of these are:
- Multiple choice questions are no good.
- Only a native or near-native speaker can get 100 % in a test.
- There should be no (intentional) language mistakes in a test.
- A test can only be used once, because test-takers may have heard the questions from a previous test-taker.
- Extended writing is difficult to mark.

This list is a fairly random collection of comments about tests from teachers. None of them is true all the time. Trying to bear them all in mind only serves to make the life of a test setter unnecessarily difficult. With testing, as in teaching, most things are relative. Let us dispel the myths one by one.

### Multiple choice questions are no good

It is in fact the case that bad multiple choice questions are no good, as are all bad test questions. Teachers are quick to criticise multiple choice questions, as many of the ones they see may have been badly set and many are not test questions at all. Good multiple choice questions reflect certain cognitive processes which occur while performing certain language activities, in particular reading and listening, and this makes them suitable for some tests. Good multiple choice questions take a great deal of time and effort to set, a lot of editing and a fair amount of wasted material, but once they have been set, they can be very good. The time invested in the setting of good multiple-choice questions is returned with interest when they come to be marked. Any good test requires a certain amount of time for its setting and marking, and it is a question of how this time is divided up rather than how long or short it is. A good test can be quick to set and require more time to mark or vice versa, but a test format which is quick to set and mark, or requires a long time for both is highly unlikely to produce usable results.

### Only a native or near-native speaker can get 100 % in a test

Although relatively little is said in standard works on language testing about level, it must be fairly self-evident that any test which sets out to test whether learners have reached a certain level of language can achieve a result of 100 %.

Setting a test in which none of the takers can possibly get full points may make it invalid per se. There is a natural reluctance among teachers to give 100% for a test and many institutions have up to now demanded a bell curve of results. However the ever-increasing influence of the CEFR should lead to more understanding of the concept of level and cut-off scores and do away with this particular misapprehension.

▶ Chapter 5

### There should be no (intentional) language mistakes in a test

This again depends very much on what you are testing. It may be true that learners should not be exposed to mistakes more than is necessary during the learning process but there are many things which apply to learning, but do not apply to the same extent to testing.

How would you test, for instance, proof-reading or correction skills without including mistakes to correct? Again, once you have decided what you are going to test, you can then decide how to do this, and if a sensible way involves choosing a correct alternative from a range including incorrect language, you may decide to do this.

### A test can only be used once, because test-takers may have heard the questions from a previous test-taker

This will depend very much on the type of test, the culture of test-taking prevalent, questions of fairness as well as the main issue of reliability. Evidence of reliability can only be provided if a test is taken more than once and the results from different sessions compared. This does not mean that the test has to be identical but that different versions must be consistent with each other and equally valid in terms of the specifications. This addresses the difference between a test format and specific test questions.

### Extended writing is difficult to mark

A test involving extended writing is no more difficult to mark than any other test, although it may take longer. If you know what you are testing and the marking criteria you are using, it will not be difficult. This is a question of criteria. With good criteria which are easy to apply, extended writing can be very easy to mark. It is (again) a question of knowing what you are testing and why.

▶ Chapter 13

■ Before you actually start the process of test setting, i.e. designing a test format and thinking of topics, tasks, tests and questions, it is important that you give some thought to what your own idea of BE is and how you see this in terms of the test you want to or have been asked to set. Assuming you want to do the job of test setting well, you will want not only to be satisfied yourself with the result but also to convince others that the test you have set is a good one, both in the short term and possibly also in the long term. Amongst other things, you will need arguments for the validity and reliability of your test, as well as its objective and administrative practicability. ■

**To think about:**

- What are your prejudices and preconceptions about tests?
- What do you want or need to know?

CD-ROM

*"The universe of discourse postulated for the learner's control is too vast ever to be completely tested. The test samples the situations, the items and the features which the learner should have grasped."*

Davies, Alan (1977): The Construction of Language Tests, in: Testing and Experimental Methods. The Edinburgh Course in Applied Linguistics, Vol. 4, OUP, p. 43

# 2 Types of tests

## Using tests for the right purposes

- What different types of test are there?
- What are the different issues involved?

This chapter will deal with different types of test which are generally used and which BE trainers may be asked or want to set. The principles of testing which are described in this book generally apply to all types of test. The differences between different test types lie mainly in the types of items in the test, their weighting and scoring and what the results are used for. These differences mean that test setters have to consider slightly different issues depending on the type of test they are setting. The following test types will be dealt with in this chapter as these are probably the most common types BE trainers usually have to set:

- Placement
- Diagnostic
- Proficiency
- Achievement
- Progress

All these types of test overlap with each other in certain ways, perhaps inevitably, as they all (should) share the same features of good tests. The differences generally lie in the focus of the particular type of test, the questions set and the use to which the test results are put.

## 1 Placement tests

**What is it?**

The aim of *placement tests* is to put learners into an appropriate course for their level.

The word *test* as applied to placement and diagnostic tests is not really appropriate as it tends to make the test-takers nervous. Learners have even been known to talk about "passing" or "failing" a placement test. Depending on their attitudes towards the placement procedure, it may be better to call the test a *procedure*, a *survey* or if it is a test of speaking, an *interview*. In this chapter we will use the term *test* as it is the most suitable one for BE trainers to use.

However it is worth considering carefully how you wish to describe the placement procedure to the test-takers. It is certainly very important, as it is with all tests, to explain the purpose to learners. This applies particularly if you are using one test for a large group with different levels of English and possibly quite different attitudes to learning English and BE courses.

The aim of placement tests is to put learners into an appropriate course for their level. What is appropriate and what we mean by level will depend on external factors. For example, the course(s) may already exist and be running, so that the learners join them once they have the results of the placement test, or the intention may be to design a course for the learner(s) based on the results of the test. In the second case, there will be overlaps with diagnostic testing as the results of the placement test will be used to make a statement about the learner's strengths and weaknesses and design a course to deal with these.

A placement test can be used when courses already exist, or to set up a course or several courses for a group of learners. There may be no preconceived ideas about the levels of the courses, so that these will be decided on after the placement test. For example, the test results may indicate that all the learners are (more or less) at the same level and can all take part in the same course. This means that division into two or more groups may simply be a case of numbers. Alternatively, a large group may be divided into two or more smaller groups and take part in different courses on the basis of their test results.

Another possibility is that the learners may be placed in courses which have been or are to be designed using an existing concept, such as the CEFR. The construct and content of the placement test will depend on this, i.e. what concept lies behind the course design. Another possibility is that the test is used to exclude learners at both ends of the scale. Those with very high scores may be considered not to need (more) English lessons and/or those with low scores not to qualify for English classes.

If learners are divided into different levels, but have to take the same course in the same and with the same end-of-course assessment, which is the case at many universities, think about whether placement really makes sense. It may make more sense to have mixed-level groups and a teaching approach to deal with this, or use tutorials, remedial teaching or more lessons.

All these issues should be considered as they make up the specifications of the test. You should also consider how binding the results of the test are, i.e. how easy it will be for learners to change courses or for new learners to join courses at a later date.

If you are asked to set a placement test, find out as much as you can beforehand about the learners. The following information will be useful:

- What experience do they have of learning and using BE? (How long? How long ago? Where? How often?)
- How much time is available for them to do the test?

- Will they be supervised while they are doing it?
- What constraints are there on the placement after the test? How many groups should there be? How big should they be?

Then decide:

- Should the test be only a written test or should there be a speaking component?
- How will this be conducted?
- By how many people will this be conducted?
- How many rooms will you have?
- How much time will there be for it?

## Use the learners' own judgement

If possible, use self-assessment to find out as much as you can about the learners' level of English. Make it clear to them that they will not gain by assessing themselves at a lower (or higher, but this is less often the case) level than they are really at.

Self-assessment (and possibly an interview) is also important to make sure that learners do not put themselves into beginners' or low level courses because they think they need to "start again". It is a common misconception among learners whose experience of learning English is a long time ago to assume that they have forgotten everything and need to go over all the basics again. In most cases, they are bored and frustrated in low-level courses, and those learners really at the low level are frustrated by the speed at which these learners (who really belong in a refresher course) pick everything up. Although BE is not usually taught to complete beginners, this principle applies to the lowest level of any course. Make sure that only those at the lowest level you want are really placed in the course.

## Define the minimum level for the lowest level course

This can be done best by defining this level as clearly as possible and setting a part of the test to test it. You may want to ask simple grammar questions to find out what learners know and then test whether they can actually use these structures in an oral test.

If learners at different levels are all taking one test, make it clear that they should not guess wildly at the answers. A certain amount of intelligent guessing is acceptable, but the results will be more useful if test-takers do not guess completely randomly. One way of dealing with this is to ask learners not to guess randomly, but rather to indicate if they are not quite sure of an answer and are guessing intelligently. They could also be asked to eliminate those answers which they are sure are wrong and mark all answers which they think could be right. If they have absolutely no idea, they should not answer the question at all.

The more detail you can get about a learner's level, the better, so if there is time, discuss (particularly wrong) answers with learners and try to find out why they answered in the way they did. One method sometimes used to decide on the efficiency and suitability of a test is to ask learners to talk through the test with test-setters as they do it, describing and reflecting on their thought process and their reasons for choosing or giving the answers. This of course is very time-consuming but can provide helpful insights into the test and its usefulness.

If the learners are all at a similar level, whether low or high, then most of the items in your test should be at this level. You will however need a fairly large number of items to differentiate between the learners.

### Don't reinvent the wheel

If you are planning to use a course book, see if there is a placement test available for the book or series. These tests are designed to see how much of the material in the course learners already know. They are often available as online tests or as downloads from the Internet. While they may not be useful for assigning learners to other courses or a particular CEFR level, they can save you a lot of work and usually provide reliable results for the particular course concerned.

### Decide on the course design

If you have no clear idea about the levels of the learners and have to use the results of the test to set up your own courses, you will need to have thought about of what you intend to teach, what the aim of the course is and how much time you will have to reach it. If you are designing the course(s) yourself, you will have to decide on the focus they will have and set parts of the placement test with this in mind. Although most placement tests are grammar-based, results showing learners' knowledge of grammar may not always be useful if the course is based on communication skills, such as giving presentations or telephoning.

▶ Chapters 7–12

### Decide how much money you want to invest

There are excellent commercial placement tests on the market, including computer-adapted ones, but these are often too expensive for companies or individual trainers to want to make the investment.

▶ Chapter 6

If you do decide to spend money on a commercially available placement test, read the specifications and follow the guidelines for choosing an external exam. There are also many so-called "placement tests" available on the market and particularly the Internet which claim to place learners reliably but in fact do nothing of the sort. BE trainers can often set better tests themselves.

Many generally available placement tests are tests of grammar and vocabulary. These test knowledge rather than ability. Although the results of these may be useful at low levels, they will not be as helpful for assigning learners to higher levels as the answers given say nothing about other skills, such as writing or listening. Tests which test more than basic vocabulary will also not be useful if the learners simply happen not to know the items tested.

### Decide on the questions

Placement tests usually consist of graded questions, either written or spoken. If a test-taker answers a question correctly, he/she goes on to the next one. This continues until the questions become too difficult. Sometimes separate tests are provided and the test-taker is only given the next test if he/she gets a high score. Using separate tests is usually a better way of placing learners. As a result of the self-assessment, learners get the test which seems at their level. If they get all or most of the questions right, they go on to the next level. If they get most or all of the questions wrong, they take the test for the level below. This is the basic principle behind adaptive tests.

If you do not know much about the learning background of the test-takers, it is best to use question types which they can understand easily, such as multiple-choice. This will mean that you do not have to spend time explaining how to answer the questions properly. Make sure that the instructions are clear to the test-takers. They can be in their own language if necessary.

### Find out how long the test can take

The number and type of questions will depend on how much time is available. In general an hour is enough for fairly reliable placement. In terms of multiple-choice questions, this will mean about 100 questions in all. The time a test-taker needs to do the test will also provide an indication of level, so it is worth making a note of how long each test-taker takes to do the test. This is of particular interest if most of the questions are answered correctly in a short time as it will indicate a high level. However, test-takers should not be asked to do a written placement test as quickly as possible. A written test of vocabulary and grammatical accuracy will provide results about someone's level of accuracy in written language. There is no reason why the questions should not be answered carefully and with some consideration.

If you want to test reading or productive skills, then the speed at which a text can be read or answers to questions given will also be important. For these tests, a time limit should be set and/or test takers asked to answer as many questions as they can as quickly as possible. You should have more questions than you probably need so that the total number of questions which each test-taker answers can also be used as an indication for placement.

### Decide what everyone must know or be able to do before starting the course

Try and test this. Do not base the results on just one task or question, but have two or three items or questions testing, for instance, being able to talk about work, answer the telephone, talk or write about the past appropriately or use conditionals (e.g. for negotiating). Decide how much of each is necessary to join the course at all, or to join each level of the course.

### Make sure that there are some questions which all test-takers can answer

If everyone gets the answer to a question right, it does not mean that the question or the test is too easy. This type of item is sometimes known as an *anchor item*, as it provides you with a starting point, in this case for placement. However if everyone can answer all the questions, then the test is too easy for the group and you will not be able to do anything useful with the results. Likewise the test is too difficult for the group if everyone can only answer a few questions, especially if these are all different ones. Remember: learners who are complete beginners do not need to take a placement test! Ideally all or most learners should be able to answer the easy questions, with the number decreasing as the questions become more difficult.

### Test more than just one skill

For instance, if you have a set of multiple-choice grammar questions, ask learners to write a page about themselves as well. If you have discrete vocabulary questions, have a reading comprehension test where knowledge of vocabulary in context and reading strategies can be tested.

### Interview learners in English

Have a range of questions ranging from simple and straightforward to more complex. Decide on one or two simple criteria and place learners on the basis of these. These should be connected with what you want to do in the course(s).

### Use different criteria for testing different things

If you do a grammar test, use criteria other than grammatical accuracy (range of vocabulary, creativity, fluency, pronunciation) in the writing or speaking test.

## Discuss the results with the learners

See if your placement coincides with theirs. The results of a placement test should also be checked after courses have started to see whether learners are in the right courses and how much adjustment has had to be made within the courses. This is easiest to do if you are the person running the courses. Otherwise you will need to talk to the trainers concerned. Compare the results of a placement test with end-of-course results to see if the placement matches these, bearing in mind the differences between placement and achievement tests.

## A placement test needs to be practicable

This usually means quick to run and easy and quick to mark. However a valid and reliable test will take up to an hour to do, so this should be made clear to whoever has commissioned you to set the test. It is best if everyone takes the test at the same time under the same conditions. Because of the pre- and misconceptions about placement tests, test-takers often behave in such a way that the results are falsified, i.e. by asking colleagues or looking up the answers to questions. Unless you can be absolutely sure that they will do the test under exam conditions and treat it confidentially, it is better to insist that the test is taken in supervised groups.

## Decide what to do with the results

Placement tests designed to sort out those test-takers who are not at a certain level (so-called knock-out tests) overlap in their intentions with achievement tests and are slightly different from standard placement tests in that test-takers can in a sense "fail" the test if they do not get the result they need to join a course. For this reason, these tests have to be set slightly differently from standard placement tests. They do however have the advantage that test-takers will try to do their best to get into a course at all.

*All Harold's aptitude test showed is that he had an aptitude for taking tests.*

If learners are divided into different levels but have to take the same course in the same time and with the same end-of-course assessment, think about whether placement really makes sense. It may make more sense to have mixed-level groups and a teaching approach to deal with this, or use tutorials, remedial teaching or more lessons.

Make sure you can explain your results to the test-takers. This means looking at more than just the numbers of points they have gained in the test, but also where they have got these points, the points they have not got and how the different parts of the test refer to each other.

### To think about

If you have to divide a larger group into two, don't always do it down the middle. See how the results of the test match among test-takers. It may be more effective in the long run to have two groups of slightly different sizes.

## 2 Diagnostic tests

If more than one skill is tested, the results of placement tests can sometimes be used for diagnostic purposes. However the purpose of a diagnostic test is not only to decide on a learner's level, but also to provide the learner with teaching suitable for his/her needs. Many of the issues described above will have to be considered, but you will definitely need more than one separate test if you are going to use the results to determine the content of a course.

The aim of diagnostic tests is to find out where learners' strengths and weaknesses are. This should be done for a purpose, for example in order to design a course, to advise a learner what course to take or what he/she should concentrate on learning or revising.

It is important to bear in mind that many strengths and weaknesses will be relative and that a consistent picture for a very small group may not be possible. If your course is to be designed for a very small group and intended to teach those areas where learners are weak, then you will need to identify the skill, find tasks at one level to test these and compare the results.

In order to have comparable and valid results, you will need to invest some time in making sure that the tasks are at a comparable level. You can use CEFR level descriptions to help you or examples of tasks from a unit of a text book. Sample papers published by exam boards may also be helpful.

Find out from the learners (or those responsible for organising courses) what the aim of the course should be. Then decide which skills and sub-skills you will need to teach. Design a test for these using one or two levels and compare the results for each learner. For example, you need to design a course on presentations and want to know whether to focus on fluency or accuracy. A written test may consist of 50 questions on general grammar and on presentation gambits (e.g. signposting language) in a multiple-choice format where the correct answer depends on knowing the exact words and structures to use. An oral interview could consist of summarising a short (simple) text freely. This oral part could be done in pairs or groups to save time, with test-takers "presenting" their text to each other. The main criterion would be fluency (you are testing accuracy in the written part), with possibly correctness of content, use of own words and accuracy playing a secondary role. However as a diagnostic test to distinguish between accuracy and fluency or between knowledge of grammar and vocabulary and the ability to speak fluently, it is better to limit the criteria to as few as possible. The more detail you have on sub-skills the easier it will be to design the course.

▶ Chapter 10

## 3 Proficiency tests

Probably the world's best known English exam has this word in its title. The Cambridge ESOL Certificate of Proficiency in English (CPE) was originally developed to ascertain whether a learner of English had the necessary skills to do a degree course at an English university. Now, many people take CPE for other reasons, but the principles of proficiency testing still apply. It is clearly not possible to test whether someone can actually study in English before they start doing it so how can it be tested? Setting a proficiency test means

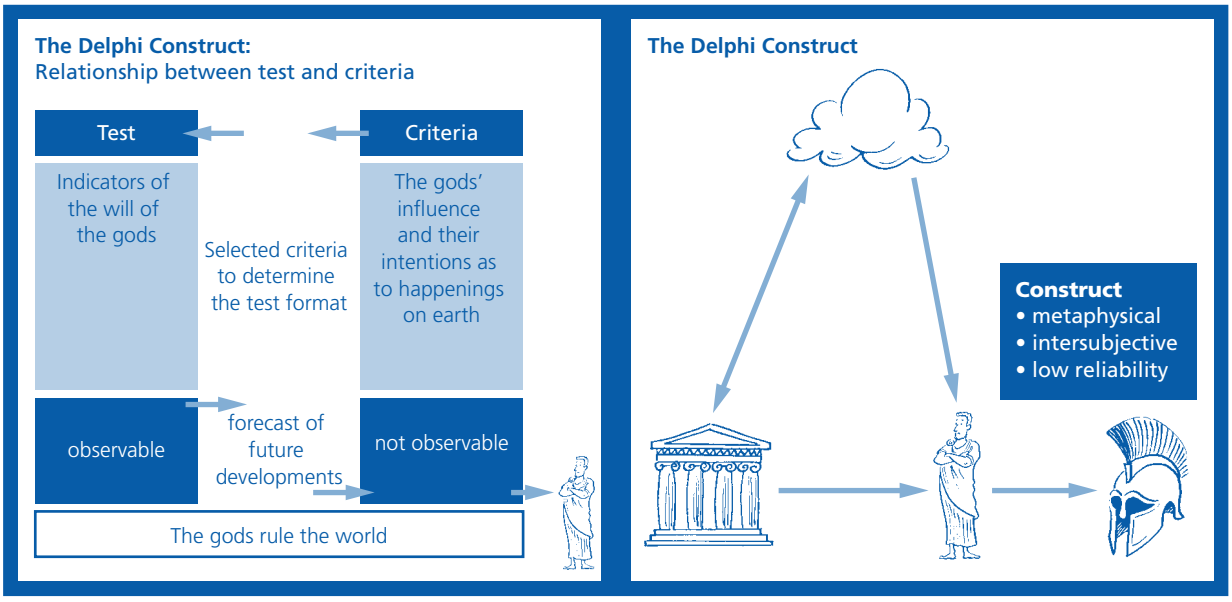
- identifying the skills, knowledge and abilities necessary for the real-life task concerned,
- creating test tasks which replicate these abilities or the skills required for them and
- using these to produce a forecast of what learners will be able to do.

A proficiency test may include tasks which test-takers have never seen before. It can be compared to a driving test in which the test-taker drives without the dual controls in the car. Performing a selection of activities gives the examiner enough information to decide whether the driver can be let loose on the road. This is what a proficiency test sets out to do, give enough information to make a decision about future performance. The main thing about setting a proficiency test is deciding what to test and making sure that it is in fact representative and provides a good forecast.

There is a certain overlap between proficiency and achievement tests if the test-taker has in fact already done the tasks concerned and is merely demonstrating for some external purpose that they can do them. This can be compared with a driver taking a driving test in another country having passed it somewhere else. However a proficiency test should involve dealing with unfamiliar situations and tasks and will involve using strategic competence rather than rote learning to pass it (if it is a good test).

Proficiency tests are probably the most challenging type of test to set and for this reason are the ones that BE trainers are least likely to have to set. It is important to understand the principles in order to identify a good proficiency test as well as to be wary of making statements about future proficiency on the basis of results from other types of test.

Using the results of tests to infer performances can be illustrated by the "Delphi construct" which also illustrates nicely the concept of construct. The Delphi construct is based on the assumption (construct) that the gods rule the world. When this assumption was prevalent, the gods indicated their intentions through the flight of birds, the state of entrails in fowl or the surface of a copper kettle. These signs were interpreted by experts (the oracle at Delphi) who used their expert observations to make forecasts of future performance. They provided these forecasts to army generals like General Gyros (shown here), who then went confidently into battle, assured of victory. Whether their forecasts came true or not, we do not know. The point is that the "test" had very high construct validity, because everyone believed in the construct, just as for many years everyone believed that knowledge of the grammatical rules of a language was a valid forecast of language proficiency.



## 4 Achievement tests

Achievement tests are probably, along with placement tests, the main type of test that BE trainers will be asked to set. Achievement testing includes end-of course testing and exam-setting as well as designing and using classroom tests. In the case of classroom tests, there is a certain overlap with progress tests depending on the use to which the test results will be put. In achievement tests, the main question is whether the test-taker has achieved the goal which was set, for instance at the beginning of the course or programme. The results of any individual test-taker will depend on a combination of two factors:

- how successful the teaching was in providing learners with the means of achieving the goal and
- how far the test-takers have learnt what they were supposed to.

This means that the results of an achievement test can also be used to assess the success of teaching, although both sides of the questions (teaching and learning) have to be taken into account.

As achievement tests test what has been taught, they should consist of familiar tasks on familiar content and can produce results of 100 %. There is no reason why all the test-takers in a group should not achieve the goals set, although this may mean that these have perhaps been set too low. This may have consequences for the design of further courses for the same group. Remember it is not the test-takers' fault if the test is too easy, and they should not be penalised for this.

The content of an achievement test can be decided on at the beginning of a course, perhaps not in the first lesson but fairly early on. Decide what you feel you and the learners can reasonably achieve in the time available, how much input you are going to give them, and how much self-study they will have to do. Make the goals clear to them and let them know throughout the course which aspects of the material will be important in the test. This is not teaching to the test, but testing what you teach. The important parts of the test are a result of the goals set and not the other way round. In this way the test will be as fair as possible and allow learners the best opportunity to perform to the highest possible level.

Identify which aspects of the goals can be tested in the time, i.e. which are representative of what has been taught and should have been learnt. Make sure you give learners as much input on the content of the test as possible. Tell them how they can achieve the highest results in the test and give them the opportunity to practise and learn what is necessary to do this.

**Example**

The aim of the course is to teach and practise negotiation skills. Students are given input in the form of vocabulary, grammar and phrases necessary for negotiations; they are given information on the form and conduct of negotiations, and are given the opportunity to practise a negotiation using a simulation. This is similar in form and content to the one used in the test. They are told what aspects they will be tested on (accuracy, fluency, task management, pronunciation, interaction etc.) and what criteria are being used for these. It may not be necessary to test them on all the aspects of negotiations they have dealt with in the course. The choice made will depend on how well they are expected to perform on these. For instance, if they are all very fluent and can be expected to achieve a high score on this criterion, scoring them on accuracy may be more useful as an incentive to concentrate on this. If you want them to use certain phrases and gambits during a meeting or negotiation, then this should be included in the criteria.

Clearly achievement tests should include tasks which are related to teaching and performance. They need not necessarily be related to the real word, if for example the aim of the course is to teach a particular grammar point. However if the aim is to practise oral skills then the test should not consist of written multiple-choice questions. Achievement tests can be more closely related to real-life or classroom tasks than proficiency tests.

An achievement test overlaps to some extent with a proficiency test as the above example shows. Having succeeded in a test of this nature, it can be forecast that test-takers will be able to conduct negotiations successfully. The differences are that test-takers (should) know before the test exactly what they have to learn or should have learnt in order to get the best results. Setting an achievement test can be used as an incentive to learners, but only if it is completely clear to them what they will have to do. Of course, all the principles of good test-setting apply just as much to achievement tests as to any other.

## 5 Progress tests

An achievement test also overlaps with a progress test, the main difference being in the size of the goal set and tested. The results of progress tests can be used to advise learners as well as to decide on what to teach in the future. Progress tests are based on short-term goals and smaller amounts of material and do not usually have such significant results for the test-takers. Basically all that has been said above about achievement tests applies to progress tests.

There are several other issues connected with types of test, many of which are described in more detail in other chapters (on the CEFR and on marking and scoring). The next section deals with one particular aspect of scoring which is especially important in connection with placement testing.

▶ Chapters 5 + 13

## 6 Norm-referenced and criterion-referenced tests

### Norm-referenced tests

Norm-referenced tests compare test-takers with one another and provide information about individual abilities or performance only in relation to other test-takers. Their main aim is to classify test-takers rather than to say anything about their knowledge or abilities. Norm-referencing can be useful in placement tests, if all test-takers are to be placed in courses and the level and content of the courses decided subsequently. They are sometimes used as knock-out tests, where only a certain number can be accepted for a course, for instance, but this may have the effect of raising or lowering standards without being able to make this transparent to stake-holders. The results of norm-referenced tests are less useful to test-takers than those of criterion-referenced tests. A test-taker only knows if he/she is at a lower/higher level than another test-taker, and this is of no great help if the test-takers do not know each other and are not able to assess each other's levels. For this reason in particular, these tests do not entirely fulfil all the requirements of good tests mentioned in Chapter 1. Norm-referenced tests start from the test-takers and the scores rather than from the view of language or the actual performance context. Basically any test can be used as a norm-referenced test if one aim of the test is to ensure a spread of scores.

▶ Chapter 1

## Criterion-referenced tests

Criterion-referenced tests are used to see how well the test-takers have learned particular knowledge and skills. It is quite possible in criterion-referenced tests for all test-takers to pass or to get full points as the tests show what test takers can do and what they know and not how their knowledge and performance compare to others.

The scores of criterion-referenced or standardised tests are easier to interpret but mean that any changes in the test imply a change in the standards involved. Using criterion-referenced tests can also lead to clusters of scores with certain groups, with perhaps all test-takers getting full scores. Whichever method of scoring you use, it must be made clear in the specifications. Anything else is unfair to test-takers. This brings us to the issue of fairness, extremely important and dealt with in Chapter 15.

▶ Chapter 15

■ Whichever type of test you set, it is essential to think about the issues described above as well as those connected with the skills and aspects of language you want to test. It is a common misapprehension to think that there is such a thing as a test of BE, but it is definitely not the case that any test can be used for any purpose. The validity and reliability of tests also depend on what they are to be used for so making sure your test is the right one for its purpose is one of the most important things to consider when setting it. ■

### To think about:

- What sort of test do you have to set?
- What are the practical issues you have to deal with?

CD-ROM

*"When it comes time to develop a test, simple recipes will not work."*

Davidson, Fred/Lynch, Brian (2002): Testcraft. A Teachers Guide to Writing and Using Language Test Specifications. Yale University Press, p. 2

# 3 Types of questions

## Closed, half-closed and open question types

- What different types of questions for tests are there?
- What are the most suitable question types for tests?
- How do you set them best?
- What are the advantages and disadvantages of each type of question?

One of the first questions usually asked about tests by someone who is going to take the test is "What sort of questions are there?" This chapter deals with the types of questions commonly found in tests, in general categorized according to form. The type of question most usually associated with tests is multiple-choice, but there are many more different question types which can be used.

The terms *question*, *task*, and *item* are all used in testing, often interchangeably but sometimes depending on the context. Although not all items are in fact *questions* (i. e. followed by a "?"), the term *exam question* or *test question* is often considered the most accessible for non-specialists. The technical term used by test experts or theoreticians is usually *item*, in particular when talking about individual (so-called discrete) items or in the term *item analysis*. *Task* is usually used when what the candidate has to do is more complex, or in reference to a set of individual questions or items which make up a whole. In general the term *question* will be used in this chapter.

Some types of question (e. g. more complex open ones used particularly to test communicative language abilities in speaking and writing) are particularly suitable for the testing of particular skills and will be referred to in the chapters connected with those skills. This chapter deals only with those types of question found in written tests.

▶ Chapters 9 + 10

# 1 Types of question

## Closed question types

Closed question types are those which only allow one correct answer. They are considered the most objective, although this is only the case if they are properly set. They are not suitable for the testing of active knowledge or of productive skills. They are often criticized, often for the wrong reasons (because they are badly set or are inauthentic) and sometimes because they make it easier for test-takers to guess the answer.

Closed question types include:
- Multiple-choice (also in Cloze tests)
- True/false (not mentioned in the text)
- Others (ordering, multi-matching)

## Half-closed question types

Half-closed question types are those which only allow a limited number of possible answers and require a certain amount of active language use on the part of the test-taker. As the questions can be answered in several ways, they cannot be considered as entirely objective. Criteria have to be agreed upon and if possible, sample answers provided. They test active knowledge and productive skills to a greater extent than closed items.

Half-closed question types include:
- Cloze tests of various types
- Gap-filling
- Transformation (sentence, word)
- Word definition
- Proof-reading

## Open question types

Open question types are those which lead to free production by the test-takers and are the type most often used in tests of speaking and writing. They are also the type most often referred to as tasks. They are considered as more authentic, they require criteria for marking and they are more time-consuming to mark. However they test active knowledge and productive skills to the best extent.

Open question types include:
- Answering questions (short answer, Wh-questions)
- Finding a title
- Note taking

# 2 Closed question types

## Multiple-choice questions

Multiple-choice questions usually consist of a stem and a number of options from which to choose the right answer. These options are made up of distractors (wrong answers) and a key (the right answer). In individual items (discrete items), the stem may be a part of a sentence, a complete sentence or a question. The most common form of multiple-choice question is that where there is only one key, i. e. only one answer is right. The principles involved in setting multiple-choice questions with more than one correct

answer are the same. These are just harder to set and require more checking. The minimum number of options for multiple-choice questions is three, although four or five are also possible. It takes more time and effort to set multiple-choice items with more options, but it makes it more difficult for test-takers to guess the right answer.

If you are thinking of setting multiple-choice items, it is best to start with three options of which one is the right answer. This will give you the best items in terms of validity. However multiple-choice questions can only be really valid if they are used to test the right things.

Multiple-choice questions are commonly used in tests of reading, grammar and vocabulary. It is in these tests that they are the most valid question types. They cannot be used to test speaking and writing as communicative skills. Although multiple-choice questions are also used in tests of listening, they make a test of listening more difficult as test-takers cannot refer back to a text (as they can in reading tests) but have to rely on their memory.


**Remember**

As multiple-choice questions are used in all sorts of places, from driving tests to quizzes in magazines and on television, they are usually familiar to test-takers and so have a high degree of interactivity, certainly in European contexts.

Multiple-choice questions are sometimes considered inauthentic, but if they are well set this need not be the case. When we use language we make choices all the time, even as native speakers of the language, and decide which alternative from several is the best to use. This applies as much to grammatical choices as to vocabulary and understanding texts (reading comprehension). This makes multiple-choice questions particularly suitable for testing the correct use of grammar. However it must not be assumed that making the right choice in a multiple-choice question means that the test-taker will make this choice when using the language actively. The results of multiple-choice tests of grammar and vocabulary, for instance in placement tests, can only be used to make statements about test-takers' knowledge and not their active skills.

Multiple-choice questions are also suitable for tests of reading if they reflect the choices that are made when reading a text. If multiple-choice questions are set with the above in mind, then they can be as authentic as other types of questions.

▶ Chapter 8

Multiple-choice questions are easy to mark if they are set well and the right answers agreed on beforehand. They are also relatively easy to set when they are used for testing grammar and vocabulary, but are more difficult to set for tests of reading.

**The most common mistakes in multiple-choice questions:**

It is these mistakes which make it extremely important to check, proof-read, edit and pre-test multiple-choice questions. Even very experienced test setters fall into many of these traps.

1 More than one answer is correct or can be considered as acceptable.

This is the most common mistake, particularly at the beginning. It can be avoided by checking the answers yourself. This is best done after a while when you have forgotten what you originally intended as the answers. It is also advisable to ask as many other people as possible to check them and do them, without giving them the correct answers.

If you change the answers and are providing a key, remember to change this as well. Similarly if you are arranging the options in alphabetical order, you may have to rearrange them.

2 The question can be answered using other skills than language, e.g. in the case of reading comprehension, without reading the text but instead by knowing a certain piece of general information.

**Example**

From a reading test where a text about the UN is used:
The Headquarters of the United Nations Organisation are in
a) Geneva
b) London
c) New York

One way of avoiding this is to remember that a test of language should not test general knowledge or knowledge of the topic.

One way of checking that the questions cannot be answered using other knowledge is to ask someone (a colleague or a learner) to answer the questions without looking at the text. If it is possible to get the right answers this way to any of the questions, then they should be changed.

3 One of the options is clearly wrong and so can be discounted. This reduces the number of options from which the test-taker has to choose.

**Example**

From a grammar test, the item is designed to test the past tense:
I _____ that film in the cinema last week. It was great!
a) have seen
b) saw
c) was seeing
d) will see

It is unlikely that test-takers at the right level will choose d). The choices will probably be spread between the first three options.

▶ Chapter 12

One way of avoiding this is to ask what choices the language user is likely to make. It is usually more difficult to set multiple-choice questions on grammar using four options as we do not usually make a mental choice between this many options. Using three options rather than four makes setting good multiple-choice questions on grammar easier.

One way of checking this is to pre-test the questions on the target group or a similar group and see if option d) in the case above is chosen at all. This can and should be done with multiple-choice questions with three options too! However it is often easy to spot this sort of mistake: it sometimes simply involves asking someone else to look at the questions with this in mind or even just to ask yourself what you are testing. If there are not three plausible distractors, it is probably better to use only three options or to use a different type of question.

4 To test something different in one of the options is another typical mistake in setting multiple choice questions. This not only makes the test invalid, it can also confuse weaker test-takers (and is therefore unfair). Strong test-takers will discard the option. In

a test of reading, this could mean including information in a multiple-choice option which is not anywhere in the text.

**Example**

From a vocabulary test:
I think _____ lies is wrong in all situations.
a) saying
b) speaking
c) telling
d) trying

From a grammar test:
I would like to suggest _____ a few changes to the agenda before we begin.
a) making
b) take
c) to do
d) to make

This mistake can be avoided by thinking carefully about what you are testing (in this case grammar or vocabulary) and the typical mistakes learners make, i.e. what choices they make when using the language.

5 One other mistake is to have options which are wrong in themselves. This usually happens when the test setter can't think of a final option. Examples of this are options which are illogical, nonsense or have language mistakes in them which test-takers may spot and so eliminate the option.

**Example**

From a test of word form:
I think we can start our meeting as everyone is here. I have one _____ for absence from John.
a) apologetic
b) apologise
c) apology
d) sorry

From a reading comprehension test:
The company intends setting up branches in several European cities including _____ .
a) Helsinki
b) New York
c) Paris

I am astonished that they have managed to stay afloat and not _____ bankrupt.
a) goed
b) going
c) gone
d) went

There are also a few general things to be considered when setting multiple-choice questions, whatever you are testing.

**1 "All of the above" or "None of the above" as an option**

This option should be avoided, particularly when test-takers are asked to choose the "best answer". If all of the above are correct this means logically that any answer is also correct.

The option is also automatically eliminated if one of the other options is intended and chosen as the only correct answer.

**2 Order of options**
It is easiest to set multiple-choice questions by first thinking of the right answer and then deciding on the distractors. The next step is to arrange the options in a different order. There are basically two ways of doing this, either in a completely random order or in alphabetical order. This is done in all the examples given above. A completely random order is more difficult to guarantee. Using alphabetical order is extremely simple and just as random. It may mean that the right answer is in the same place in a number of consecutive questions, but this need not confuse the test-takers if the principle is explained to them beforehand. In the case of numbers, the best order is from lowest to highest and in the case of dates, chronological. Doing all these things helps those test-takers who know the answer to find it easily.

**3 Length of options**
All options should be of approximately the same length. If one is much shorter or longer, this distracts test-takers unintentionally.

**4 Stem and options**
The stem should include as much of the information as possible and this should not be repeated in the options. The options should concentrate on what is being tested. Any information which applies to all the options should be put in the stem to reduce the amount of reading for the test taker.

**5 Question or completion**
The stem can be a question or a sentence which has to be completed. It may also in some cases be a complete sentence (e.g. where a sentence with a similar meaning has to be chosen from the options). As long as the test-takers are familiar with these different question types and/or examples are given, they can be used in the way most appropriate for what is being tested.

**6 Instructions**
The instructions for multiple-choice questions should be completely clear as to how many correct answers there are, that only this number should be marked and how a wrong answer is scored. If possible and necessary, give test-takers an example.

▶ Chapter 13

Decide whether you want to use the term *best answer* or *correct/right answer*. *Best answer* is probably more appropriate for tests of grammar and vocabulary, especially where giving the wrong answer may not lead to any confusion (*best* is defined as meaning grammatically correct or best in all circumstances). For tests of reading, *correct* or *right* may be more appropriate. The decision can be left to the test setter as long as the instructions are clear to all those using the test.

## Cloze tests with multiple-choice questions

Cloze tests will be dealt with here as well as below, because multiple-choice questions can be used in conjunction with them. In a Cloze test words are deleted from a text, leaving gaps. The test-taker fills the gaps with one or more words. The gaps can be made by deleting every 5th, 6th etc. word, or by deleting particular types of words, e.g. prepositions, verbs etc. Options can be given for each word, making this type of test a closed test. A closed Cloze test is more suitable for testing vocabulary than an open Cloze test, which generally leads to too many acceptable answers.

Some particular guidelines for setting this type of test, whether you use multiple-choice questions or not, are:

- Do not place the gaps too close to each other. Leave 6 to 12 words between gaps.
- If you ask test-takers to fill each gap with only one word, do not delete parts of hyphenated words or short forms.
- Leave in a few sentences at the beginning of the text so the test-taker can get a feeling for the content. This makes the text easier to understand.
- Try to avoid deleting the first word in a sentence or deleting negatives. Taking words like adjectives or adverbs out which leave a complete coherent sentence should also be avoided.

There are many uses for and forms of multiple choice questions, examples of which can be found in course books and exams. The basic principles described above apply to all of them.

**What is it?**

The C-test is a very specific type of Cloze test. It is set by choosing specific texts and deleting the second half of every second word according to specific rules. It has extremely detailed rules for its setting and marking and will not be gone into in detail here. There is some doubt as to its validity, authenticity and interactivity and hence its usefulness in general testing contexts and in particular in BE.

## True/false/not mentioned

True/false/not mentioned questions, if they have three options, can be considered as similar to multiple-choice questions. These questions with all three options are often used in tests of reading comprehension and with two options in tests of listening comprehension. In tests with true/false/not mentioned questions, especially in those with only the first two options, it is obviously easier for test-takers to get a high score by guessing.

As with multiple-choice questions true/false/not mentioned questions cannot be used to test speaking and writing as communicative skills. However they are an authentic way of testing certain things, again depending on the choices the test-taker has to make in real-life.

True/false/not mentioned questions are not difficult to set and they are easy to mark. They are also easy to explain to test-takers.

**The most common mistakes in true/false/not mentioned questions:**

1 There are not enough items. This means that test-takers who guess randomly can get a high score. In general, if you use true/false/not mentioned questions, and in particular in tests of listening, there should be at least ten items which provide the final score, however these are divided up. Nevertheless it is clear that random guessing with this type of question is more likely to be successful.

2 There are too many false items. This is usually because these are easier to set. The ratio of true to false should be about 3:2 or 2:3.

3 There are clues to the answer in the question. This is particularly the case in tests of reading and it is here that it is particularly difficult to avoid. A very common mistake is to use more complex language in the questions than is used in the text. In fact the opposite is a better way of testing comprehension, whether reading or listening. It is unfair to penalise test-takers who understand the texts but cannot understand the questions!

**Example**

From reading or listening comprehension:
It is possible in theory to use Scottish banknotes everywhere in the United Kingdom, although some shop-owners may not want to accept them.

Scottish currency is legal tender in retail outlets in the United Kingdom.

4 Words like *always*, *never*, *none* and *all* are used in the questions. This almost always means the answer is wrong and experienced test-takers may know this.

5 Double negatives are used. Double negatives should be avoided in tests in general as they are difficult to first process mentally and then to decide on the right answer.

**Example**

From listening comprehension:
"Let's see if we can manage without a live meeting this month. To cut travel costs for all of us, I'd suggest we try teleconferencing."

The speaker does not want to avoid cutting travel costs.
*Just as bad:*
The speaker is not suggesting going without business trips.

**Remember**

Tests should not set out to trick test-takers!

## Other closed question types

The two most common closed question types are multiple-choice questions and true/false/not mentioned, which is why they have been dealt with in some detail above, but there are other closed question types (i. e. where only one answer is correct). Most of the principles described above apply to these as well.

**Matching questions** are commonly used in tests of reading and listening (matching titles and texts, pictures and headlines) but also in tests of grammar (e. g. matching parts of sentences) and vocabulary (matching synonyms, words and definitions). These questions are not particularly difficult to set but must be carefully checked if only one answer is to be accepted as correct.

**Multi-matching** is a question type usually used for grammar and vocabulary and is similar to multiple-choice gap-filling but with a limited number of choices. Test-takers are given a text with a number of gaps. The best number is between 10 and 20 gaps in one text. The gaps should be for language items which may cause difficulty. Below the text are a number of options (vocabulary, prepositions, verbs etc.) These can consist of one or more words. The choice of these will depend on what you want to test. Only these items can be chosen by the test-takers to fill the gaps. Each item fits into only one gap.

▶ Chapters 11 + 12

If the number of options is the same as the number of gaps, it will not be possible for test-takers to get only one answer wrong and lose only one point. One mistake will always lead to a second and possibly more. Also test-takers can use elimination rather than language skills and knowledge to choose the right answers, particularly towards the end of the text. This can be avoided by providing a larger number of options than gaps. If this is done, all the options must be chosen and checked carefully.

Although this type of test can be used to test grammar and vocabulary simultaneously, it is important to make sure that each option really only fits into one gap and at the same time to avoid giving too many clues to the right answer and thus making the test too easy.

Multi-matching exercises are often created for classroom use where more than one answer can be accepted, but, in general, setting them as closed questions in a test requires a certain amount of work and several checks.

## 3 Half-closed question types

These are called *half-closed* because, in many cases, one answer is expected and most likely, but others may be possible and accepted. In general these types of question are less interactive and require more explanation and/or examples for test-takers. They can usually be made into closed items by providing alternative answers for the test-taker to choose from. Both half-closed and open questions can be used as a preliminary stage in the setting of closed questions. Answers given by learners in classroom tests can be collected, analysed and used as distractors in closed questions. This may be one of their main uses. It is important not to regard half-closed questions as closed, i.e. only accept one answer. Especially if you are marking a test like this for the first time, it is important to think about all the answers you get from test-takers and why these may have been given. Also think carefully about whether you can accept them and why or why not.

**To think about**

Although in a small-scale low-stakes test where one person is marking, mistakes in questions can be corrected later, this does not justify them.

### Open cloze tests (gap-filling)

Gap-filling tests and exercises are popular among teachers. They are usually used to test one particular aspect of language, usually grammar and vocabulary and, like closed items, cannot be used to test communicative or productive skills. They are very useful and valid for teaching structural competence but less useful for vocabulary, where there are usually several possible words which can fit into a gap. Most of the work involved is choosing the right text and then identifying the gaps and the reasons for testing these aspects of language.

Although gap-filling is easy for many test-takers to understand (i.e. high on interactivity) it is fairly inauthentic, especially if the number of words to fit in the gap is specified. Depending on the choice of text, gap-filling involves both reading comprehension and making mental choices. If the test-taker has difficulty understanding the text or context, he/she will not be able to fill the gaps. On the other hand, the test-taker may be able to understand the communicative message of the text but not know the right word to fill the gap.

If you are thinking of using a gap-filling text, ask yourself the following questions:

- What am I testing?
- What answers am I prepared to accept?
- How will I mark spelling?
- Are all the gaps of equal importance in terms of language and in terms of how I mark them?
- Can a test-taker get full marks even with some wrong answers?

### Transformation

These are tests where test-takers have to change words or sentences, usually while retaining the original meaning. Transformation tests are also used for testing grammar and vocabulary, but can only be used for certain grammatical structures. Although they are not suitable for communicative tests of writing, they can be used to test correctness in writing and the use of structures in writing.

### Sentence transformation

The test-taker is given a sentence and asked to change it using either the word(s) or a sentence beginning given. The sentence given by the test-taker as an answer should be correct and mean the same as the sentence given. For example, active sentences can be transformed into passive sentences or statements made into conditional sentences.

**Example**

Representatives from both companies signed the contract last Monday.
The contract …

We didn't get that order because our prices were higher than our major competitors' prices.
If our prices …

Transformation items like this cannot be used to test all structures and should be practised before the test. There should be one possible answer (or a small number of clear and correct answers). It must also be decided how points are to be awarded, e. g. how far spelling or mistakes which do not affect the main meaning of the sentence (e. g. in the above example, "the contract were signed") are taken into account.

### Word transformation

These are questions where a key word is given and test-takers have to use a different form of the word in a given context.

**Example**

I met a _____ of Bellman & Co at the conference last week.
(REPRESENT)

The context given should be easy to understand. The more context there is for the test-taker, the more different word forms can be tested, including negative forms.

### Proof-reading/error identification

**What is it?**

The *Monitor Model* presented by Stephen Krashen suggests that conscious use of language is only available as a Monitor. This Monitor helps language learners to identify and correct their own mistakes. This means the language user must have time and be focussed on form. This only comes into force after a certain amount of acquisition.

This represents a fairly authentic task for many test-takers, especially at advanced levels. Advanced learners may for instance be asked to look at pieces of colleagues' writing and correct these. It is not usually suitable for learners at low levels (below B1) as it involves knowing, having been exposed to and being able to use correct and accurate language.

In a test there are various forms of proof-reading, making it easier or more difficult to set and mark.

Tests of proof-reading or mistake correction must be set carefully so that only the mistakes intended are corrected. There are a number of pitfalls to be avoided. Using individual sentences is one way of doing this but makes the test less authentic as the context is missing. If a text is used, it can be divided into lines with only one mistake in each line to be corrected. The test can be made more difficult by not including a mistake in every line and asking the test-taker to identify those lines where there is no mistake. The test can be made easier by identifying the mistake and asking the test-taker only to correct it, although in general this makes it very easy. Other forms of proof-reading with their advantages and disadvantages are described below.

### Identifying a superfluous word

Although this type of proof-reading can be considered as a closed test form, it is not always easy to set as there are only a certain number of realistic mistakes of this kind which learners make (e. g. *I am agree, I must to go, I tell to you*). Also it is important to check that there are no other words which could be seen as superfluous.

### Identifying a missing word

This is based on the same principles and has the same disadvantages as described above. The test-taker has to identify a missing word in each sentence or line (e. g. There was a

bit trouble in the factory yesterday/we made lot money on that deal/She got married him last week.)

### Mixed mistakes

In this form, something in each line is wrong. The mistake can be one of spelling, punctuation, grammar or vocabulary. The test-taker has to identify the mistake and correct it. This type of test can be made easier by identifying the mistake for the test-taker. The test setter has to decide which correct versions are acceptable.

Proof-reading tests are becoming less authentic as more people use programmes to check spelling, grammar and punctuation. As tests of grammar, vocabulary and polite language, however, they do have a place, depending on the level and needs of test-takers.

## 4 Open questions

Open questions are considered the easiest to set, and for small groups may be the most valid and authentic way of testing. However with larger groups and limited time for marking, it is essential to develop criteria, and even with these, there will always be an element of subjectivity in the marking of open questions.

▶ Chapter 13

Using open questions also makes it possible to test a variety of different things rather than be limited to certain structures. They make guessing unnecessary and allow test-takers to use language individually and creatively. They can test a range of different skills and levels and be more closely related to what is done in the classroom.

In open questions, candidates use their own language and are not restricted to the language provided. This raises two main issues which have to be thought about when developing marking criteria. One is spelling, which has been touched on above, and the other related issue is accuracy. As both of these involve issues of communication and getting the message across, these must be considered.

### Answering questions (short or long answers)

These are usually used for tests of reading or listening comprehension. The questions may ask for specific information or for information based on understanding of attitude or stance. The question types may vary from entirely open questions to Wh-questions with only one possible answer. If the questions are very specific, it is worth considering whether closed questions may not be a better way of testing understanding.

A short answer question requires an answer of one word or a few words. These are used in tests of reading and listening. The test-takers should know how long their answers are supposed to be and how far they can copy from the text or write down word for word what they hear without being penalised.

All the above information should also be given if longer answers are expected. The criteria must include information on the length of the answer, the accuracy of content, language and spelling in the answer as well as how far test-takers are expected to use their own language.

Other open question forms include asking test-takers to find a title for a text or take notes on a text using a list of headings.

## 5 Test or teaching aid?

There are of course many other activities and exercises which might seem suitable as test questions. These may include asking test-takers to give grammar rules or translation from one language to another.

This chapter has restricted itself to the most common test question types for several reasons.

An exercise is not a test. An exercise usually involves practising some aspect of language several times in order to learn it. A test can be seen as a snapshot at one point in time or a performance which gives a representative view of a test-taker's ability. Certain aspects of language can be practised over and over again with one form, but it is difficult to find one adequate form for testing several different aspects of language. This is why tests usually restrict themselves to a few particular types of questions, multiple-choice being one of the most common.

It is, in fact, these common types of question which are most suitable for tests. Some (again multiple-choice) are not really suitable for teaching or practising language, which relates to one of the essential differences between teaching and testing. A test is not a teaching aid. It neither aims at providing test-takers with new information nor is it intended for them to practise what they have learnt. A test is given to find out what test-takers know and can do with the language concerned. The implications of this mean that many exercise forms used in class are often unsuitable for tests and vice versa.

These types of questions generally fulfil the requirements for good tests – validity, reliability, authenticity, interactiveness, objectivity and practicability – and make it easy to adhere to these requirements.

■ It is always worth considering the range of types of test questions available as well as thinking about questions, activities and tasks you have used in teaching before deciding which to use in your test. This will involve considering what has been described and referred to in the chapters up to now as well as the aspect of language you are testing. Once you have decided on your question types, even if this means using tried and true multiple-choice questions, make sure you set them well. ■

### To think about:

- Do multiple-choice questions in material you know follow the rules for good test questions?
- Can you find more types of questions than those described here and might they be suitable as test questions?

CD-ROM

*"Designing and introducing a new test is a little like getting a new car on the road."*

McNamara, Tim (2000): Language Testing. OUP, p. 23

# 4 Specifications

## Deciding what to do

- What are specifications?
- Why do you need specifications?
- How should they be produced?

The term *specifications* may sound technical but is in fact fairly simple. Specifications are the main instrument for developing high-quality tests. The main aim of the specifications is to make the test clear to everyone, i. e. to answer as many questions as possible about the test. There are some questions which most people dealing with a language test will ask such as "How difficult is it?", "Do test-takers need to speak/write/answer questions on grammar?", "How long is it?", "When can I take it?" and "How much does it cost?" Some questions are more likely to be asked by teachers, such as "What grammatical structures are covered?", "Are there multiple-choice questions?", "What level is it intended for?", "What do I need to teach and practise?" and yet others by administrators, employers and others, such as: "How many examiners are needed?", "In what form are the results made available?" Test experts and bodies which award recognition to tests will ask about validation, procedures and instructions for test setters. Good specifications will answer all these questions, although they may not all be published in the same format or in one document.

"Test specifications: A detailed accounting of the test format and general test design which serves as a basis for providing information for candidates and test users, and for writing new versions of a test." *McNamara, Tim (2000): Language Testing. OUP, p. 138*

Another word for specifications is blueprint which may make what is meant clearer for some. If we compare setting or choosing a test with building a house, we can see that we need something to go on before we do this, i. e. a list of things to consider when we do it, guidelines on which to make our decisions. This may include who the house is intended for, architectural details, price and availability, all of which can equally well be applied to a house and a test. What information you need will depend on your interest in the house/test. Builders or town planners will need different information from that necessary for those who are going to live in it. All this applies equally to tests.

## 1 Writing test specifications for BE tests

Writing test specifications is similar to designing a BE course, something BE trainers should be familiar with. In general a BE trainer sets objectives for the whole course before starting on the first lesson and tries to fit all the individual lessons into an overall course design. Each lesson is designed with the entire course in mind, even if not all the individual lessons are planned in detail at the beginning of the course. The course description allows a company to decide whether to run the course and allows learners to decide whether to do the course or not by giving them information on the content and methods. Similarly, test specifications describe the test design and format and allow individual versions of the test to be designed on that basis. Test specifications not only make it possible to design a test and write further versions of it, they also make the whole process of test design and setting much easier. This does not mean that it is easy! As two experts say:

> "… we believe that the step from specification to actual test-development (or writing) is the hardest of the whole test-making process and requires the greatest combination of precision and artistry".
>
> *Carrol, B. J./Hall, P. J. (1985): Make Your Own Language Tests. Pergamon*

### Why do I need specifications?

**Remember**

The more details you agree on, the more reliable your tests will be!

Writing specifications, even if you are working on your own, will help to identify and address some of the problems of testing. You may however need to provide information to someone else or you may be working with a team of colleagues who are all setting tests in parallel groups. This is sometimes the case in universities and further education colleges. In these cases, specifications provide a set of standards which everyone works to. This means that learners are less likely to complain that one test is easier than another or one teacher marks more or less strictly. Heads of languages at universities often have the task of writing specifications. If they are lucky they work with a team to agree on and then stick to the specifications.

### Which information should specifications contain?

Specifications are not written only for test constructors. Test specifications should contain as much information about the test as is necessary for all those using the test. Test specifications may be extremely extensive, i. e. include all the information about the test and all the documents relating to it, such as grading sheets, criteria, sample items (without of course supplying the actual questions for the test to be taken) or fairly short if they are only, for example, for decision makers.

Specifications are sometimes divided up, according to who will be using them, into the following components:

- Instructions for test setters (sources of texts, length of texts, number of items, length of oral exam, etc., as well as topics, vocabulary, grammar, sample questions etc.)
- Instructions for examiners and scorers (length of exam, scoring criteria, marking sheets, samples or videos of candidate performance etc.)
- Information for candidates (similar to that for test constructors, but with sample questions and tests as the focus)
- Information on how the test was developed (details of constructors, editing and pre-testing, etc.)
- Information on who the test is for, what it tests, what the results can be used for and where it is recognized.

You may have separate specifications for different groups or leave some parts out altogether. The main groups to concentrate on are probably the test setters (which may be only you) and the test-takers. The specifications for these groups are in fact fairly similar, the main difference being that the information for test-takers is descriptive and that for test setters gives them instructions.

In general, specifications should include as much information as possible. How much this is will depend on the circumstances and what the specifications are to be used for. It is unlikely that any one person will be interested in the entire specifications, so the first question to ask is the following.

## Who is interested in this test and what do they need to know about it?

Whether or not you actually write formal specifications for the test you are setting will depend on you. You may be the only person really involved in the setting of the test (don't forget that this is not a good idea at all) but it will still be essential for you to think about the questions involved. If you have to provide any sort of documentation for your test, e.g. for colleagues, a boss or the learners themselves, then thinking about the issues involved in writing specifications will help.

Specifications, however short, are used to make decisions about the validity, reliability, authenticity, interactiveness, practicability and impact of the test. Even if the terms are not familiar to everyone, they will all be interested in these things.

Learners want to know that the results they get will reflect their real performance, ability or knowledge (validity) and that they will not be treated differently from their fellow test-takers (reliability). Employers want to know how what is tested relates to the real life of business (authenticity). Test-takers want to know what they will have to do in the test and to prepare for this (interactiveness), those in charge of resources will want to know how long the test takes and how much it costs (practicability) and test-takers also want to know if the results of the tests can be useful to them elsewhere (impact). All this information forms part of the specifications.

## What should I include?

What follows is a fairly comprehensive list of things which are often included in specifications. I will start with the things which you definitely need to think about, whether or not you actually put your decisions into writing. The main things which should be included in specifications are:

- Title
- Target group

- What you are testing
- Test content
- Task types
- Scoring

**Title:** This may be a good way to start as it will help to focus on what you are doing. The title should be as concise as possible but contain the essential information.
Some examples are

- Placement test for BE courses starting with XY (e.g. Title of Book Unit 6)
- Admissions test to BE courses at B1 and above for staff of company XY
- End of course test for learners in BE course "Presentation skills"

**Target group:** Who is the test intended for? This will involve describing the target group as closely as possible as well as deciding whether the test can be used for other groups apart from the one you are thinking of. The following can be considered: age, educational background, business background, qualifications (business and language), possibly even gender and disabilities as well as anything else you can think of which defines your target group. If, for example you are writing a test for trainees aged 16 to 19 at a company, you may make different decisions about the content than if the test is intended for business executives with at least 10 years' working experience. On the other hand, you may not want to restrict yourself in this way and decide to set a test which can be used for both groups, for instance for placement purposes.

**What you are testing:** This should be a brief general description of what is to be tested and may well be the same as the learning objectives for a course. You might also like to say why you are testing this, if it is not obvious. Here, as in all other parts of the specifications, you should use formulations which are easy for those concerned to understand. Some parts may also have to be formulated in another language than English.

It may be easiest to write a general description after you have made some other decisions on detail and can decide who needs to know what. Most probably, an employer will not be interested in the details of such things as how long the texts in a reading comprehension test are or which grammatical structures have to be mastered, but these decisions are crucially important for the test setter. It is likely that learners will want to know things like this. Once you have defined what you are testing, it is sometimes simplest to produce a sample test for learners so they can see exactly what the "real" test will be like. All reputable exam boards do this. For the individual BE teacher who is setting a test, setting a sample test can also help to practise test-setting before you go "live" with candidates and can act as a form of trialling for the test format.

*This test doesn't understand me!*

▶ Chapter 1

▶ Chapter 5

What you are testing involves the "big" question of what you mean by BE (discussed in Chapter 1) as well as the "smaller" questions of skills, level and language elements. As far as level is concerned, if you use the CEFR level descriptions (A1 to C2), you will need to know what these mean and which of the descriptors you want to and can test. It is rare that you will be testing one skill in isolation. A test of reading comprehension may involve some writing, a test of listening comprehension may involve some reading; an oral test will most likely involve listening and so on. Identify all the skills involved and describe them in the specifications.

It is essential at this stage to first decide on what you want to test and then whether it can be tested. It is important that you are aware of what you can and cannot test and do not make claims about the test which it cannot fulfil.

An example of this is testing whether learners can produce full and grammatically correct sentences in a test of spoken communication. It is rare that any speakers do this when communicating (which is the point of speaking, after all) so that it makes far more sense to develop ways of testing whether test-takers can get a message across, interrupt politely or link contributions in a discussion.

**Remember**

Not all aspects of language and language use described in the CEFR can be tested in a practicable way. In some cases, particularly with productive skills, there may be a case for other forms of assessment (self-assessment, portfolio or continuous assessment) rather than a test.

Another example is testing pronunciation using the phonetic alphabet. This may seem an unusual way of testing pronunciation, but is in fact not uncommon. A test of the phonetic alphabet will only test how well learners know the phonetic alphabet, but not whether they can actually pronounce words correctly. Asking them to actually say the words in sentences is a far better indicator of their pronunciation, but may be rejected for practical reasons (as many tests of speaking are).

**Test Content:** Define where the material for the test is to come from. Is it print material, taken from books, magazines, newspapers or the Internet? Is it adapted in any way or is it specially written for the purpose? A placement test of the type mentioned above to place learners in a course using a particular book can obviously use material from the book to see whether the learners can cope with this. An admissions test will use different material from an end-of course test. If you want to test whether the learners can transfer what they have learnt in a course to other similar tasks, the areas these will be taken from should be described. This can mean describing the sources of the material, but can also include listing the topics test-takers should be familiar with and the vocabulary involved.

**Remember**

The CEFR has lists of topics relating to the occupational domain which may be useful.

The description of the language in the specifications will of course depend on the skill concerned. This can include functions (agreeing, greeting, giving instructions, explaining), sub-skills (summarizing, maintaining a conversation, identifying the main point of a text), understanding (native speaker or non-native speaker language) as well as grammar and vocabulary.

**Remember**

A test should not contain tricks. It is not a competition between the teacher/institution and the test-takers, but rather an opportunity for both.

Teachers often feel that by describing these things in detail and putting them down on paper, they are giving too much away. Why they think this way is not clear. A test is designed for a purpose and in order to fulfil this purpose, it should allow the test-takers the chance to get the best possible results (without cheating of course). For a test to be fair, test-takers should know exactly what is expected of them.

This means that learners will know what they will encounter in the test in terms of language.

**Task types:** Test-takers should also know what they are expected to do, e. g.

- answer multiple-choice questions (how many, how many choices, what order are they in),
- write answers (how long should the answers be, how many mistakes can they make),
- fill in gaps (how many words for each gap, whether spelling counts),
- answer questions on texts they read or hear (how many questions per text, are they in the same order as in the text),
- speak in an oral exam (how much they have to say).

All this means describing the test format in as much detail as possible.

**Scoring:** Learners should also know how they will be marked. Which answers are accepted as correct? What if they give an answer you hadn't thought of? Is there only one correct answer? How much do grammar mistakes count? What about spelling? What role does pronunciation play? What if they don't know the word for something?

This means defining the criteria for marking. Of probably the most importance is the score and the final results (pass or fail): How many points are needed to pass the test? What grades can they get? How are these related to the scores? How are the questions weighted? Is there just one point for each question? Is grammar more important than spelling? Who will mark the exams? What happens if they fail? Can they take the test again? How many times?

▶ Chapter 13

## 2 Designing specifications

There are many ways of designing specifications for test setters and learners and there is no standard format for them. One way of writing specifications is to describe the objectives of the test, and then the question types, then the answer types expected and give a sample question and answer as well as any other information.

Another way is to describe the whole test (possibly in a table), including number of questions, number of points, timing, instructions, description of questions, criteria.

**Remember**

A test does not exist in a vacuum but must reflect a view of language, such as the one expressed in the CEFR in order to be valid for its purpose.

Test specifications can also be designed using an FAQ format. Think which questions you would ask about the test if you were a learner or a trainer or an employer and see if you can answer them. This may be a good way to start.

It is often interesting and worthwhile if you have the time, to attempt to formulate specifications for an existing test which does not have any. The easier it is to do this, the better the test is.

## 3 Do the specifications work?

You may find that you develop specifications and test items alongside each other. There is basically nothing wrong with this, as long as you are aware of the possible shortcomings of the approach. It is clearly impossible to write specifications without seeing whether they work in practice, but it is equally difficult to write a test without any specifications, i.e. without knowing why you are writing it. Items and tests written using preliminary specifications can be tried out on learners first (and their comments taken into account) before the specifications are published or made available to others in any form.

### Sample questions and papers

At least one set of sample questions should be provided. This may be an earlier version of the test which has been pre-tested and possibly discarded for various reasons. Sample papers should not be inferior in quality to "real" tests, but need not be of equally high quality as answers and results can be explained and discussed in class. Their purpose is to give learners a clear picture of the format and level of the test they are going to take. Samples papers are sometimes at a slightly higher level than actual tests so that test-takers do not have a nasty surprise in the test itself.

### Pre-testing

Pre-testing is one of the most important stages in test development. Unfortunately it is the one step which is least practicable and least often done outside large testing institutions. It is actually the most interesting stage of test development, as it is possible to see whether your test really works by trying it out on the right groups of learners. The disadvantage is of course that you cannot use the texts and questions in a "real" test with the same learners. But the positive side is that once you have decided on your format, you can try it out on learners, get their comments and use the results to improve the test. Then you can continue to use the items (improved of course) as sample papers to give learners and anyone else interested information about the test.

**To think about**

Learners are usually willing to comment on tests and often appreciate being taken seriously. As tests are being set for them, it seems only fair to take their views into account!

### Keeping to the specifications

It is always tempting to ignore the specifications when writing new versions of a test, but if you have gone to the trouble of producing specifications, then you should stick to them. Always keep them at hand and consult them as much as possible. If you ask others to check, comment on or edit your test, then give them the specifications to work with too. This applies however short your specifications are, even if they consist (in an extreme case) of only a general description of the test itself.

### Revising and editing specifications

Specifications are of course essential but once you have got them (remember they need not be long), it is best to see how they work over a few testing sessions so that you have a fairly representative sample of tests which you have developed from them. After this, you may wish to add to your specifications and possibly to change them. Minor improvements can always be made but major improvements may mean informing everyone concerned about changes.

### Setting standards in teaching and testing

There are standards for almost everything today and teaching, and testing are no exception. Writing specifications is no different to setting standards in testing and will provide yourself and anyone else with something on which to assess your performance as a teacher. The specifications you write will describe what you realistically expect learners to be able to do and how you test whether they can do it. Specifications therefore connect teaching and testing very closely, and the more specific your teaching context is, the more specific your testing specifications can be. However specific they are, i. e. even if you are only setting one test for a class, they will still help you to evaluate your test and also your teaching.

## 4 Some examples of specifications

These are just two examples of specifications which may be of interest to BE teachers and which serve to show how different test specifications can be, without necessarily being of inferior or superior quality.

#### Cambridge ESOL exams

www.cambridgeesol.org

Each handbook includes the full specification for the exam, guides to each of the papers or components, full sample papers, audio files for listening paper, and detailed information on how the exam is assessed (quoted from the website).

www.lccieb-germany.com

**The LCCI Practical Business English Certificate Infopack** available on the website contains a sample paper, answer sheet and syllabus (a list of language items).

■ This chapter is not intended to be theoretical or to make you feel that you have to do a lot of extra work before you can set any test at all. You may never need to write specifications and only to produce and show a test, but in order to be as fair as possible to the candidates it is essential to have at least thought about some of the things described here. ■

CD-ROM

### To think about:

- What aspects of specifications will be most useful to you?
- Which parts of the specifications might you have to make available to others?

*"A further intensification of language learning (…) is necessary in the interests of greater mobility, more effective international communication combined with respect for identity and cultural diversity, better access to information, more intensive personal interaction, improved working relations and a deeper mutual understanding."*

Common European Framework of Reference for Languages: Learning, teaching, assessment (2001). CUP, p. 5

# 5 The CEFR

## A shared point of reference

- How to use the CEFR
- What does the CEFR say about assessment?
- What do the levels mean and where does BE start?
- Which of the skills described are useful or necessary in BE?
- How can they be used to produce good tests?
- Does the CEFR need adaptation to BE and is this allowed?

### 1 The CEFR's relevance for Business English

*The Common European Framework of Reference for Languages: Learning, teaching, assessment* (to give it its full title) was published in 2001 after many years were spent on its development. It has had an enormous influence on language teaching and learning in Europe and far beyond, and is certainly important for test setters because it puts any test in a much wider context.

The main significance of the CEFR for BE trainers and test setters is that it is becoming more widely-known among personnel managers and those responsible for in-company training. Universities are also using the levels as a benchmark for entrance requirements and end-of-term and end-of-course language exams. Although it may not be possible to eradicate all the misconceptions which exist about the CEFR, it is certainly helpful to know what the levels actually mean and to be able to transfer them into a business context.

This is particularly so because the level descriptions are based on skills rather than knowledge and in a business context, it is usually language skills which are far more important than language knowledge, for instance of grammar. The CEFR provides a

▶ **Further information**

**What is it?**

What is the best abbreviation for The Common European Framework of Reference for Languages? Although CEF is commonly used, the best abbreviation is probably CEFR as this includes "R" for the word *reference*, which is what it is intended to be, rather than a prescriptive framework.

**Did you know?**

The CEFR is used as a basis for language education in one way or another in all countries of the EU and the Council of Europe. Outside Europe it is used in New Zealand, Japan and Chile. At present there is a major project under way to attempt to align CEFR levels with those of the ACTFL (American Council on the Teaching of Foreign Languages).

construct on which a test can be based and so can be used to maintain the validity and reliability of a test. It is therefore relevant, not only for deciding on the level of a test or of the test-takers, but also for designing specifications and for setting standards, whether in a general language or a BE context.

Language learning, teaching and testing is usually of skills, and all BE trainers will focus their training on the four skills of reading, writing, speaking and listening to different degrees of intensity. Although tests will be probably of one or more of these skills, setting and marking these tests will involve thinking more about the sub-skills involved, and this is where the CEFR can prove extremely useful for BE trainers who want to set tests.

What many do not know is that there are 58 descriptive scales in the CEFR. Many of these are of skills which are highly relevant in business contexts. If, for example, a learner wants to be able to take part in or conduct interviews (job interviews, customer interviews), there are descriptors for this skill in the CEFR, from A1 ("Can reply in an interview to simple direct questions spoken very slowly and clearly in direct idiomatic speech about personal details") to C1 ("Can participate fully in an interview, as either interviewer or interviewee, expanding and developing the point being discussed fluently without any support and handling interjections well", CEFR, p. 82). These, along with other relevant descriptors, such as phonological control, identifying clues and others, can form a starting point for a test of speaking.

As its title suggests, the CEFR was never intended to be prescriptive. However it has become difficult if not impossible to ignore its influence on the field of language teaching and business. Many people outside the field of language teaching, including personnel managers and those responsible for language training in companies, have heard that levels of language are now called by letter and number combinations like A1 and B2, even if they don't know where these levels begin and end. They may not know much more about the CEFR.

**Did you know?**

The Council of Europe is not the European Union and is not linked to the European Union. Its work can be seen as cultural rather than what is generally understood as political or economic. For more information see www.coe.int.

For a professional BE trainer, it is essential to know more than this for all sorts of reasons, including assessment. Although the word "assessment" is a part of the title of the CEFR, only one chapter of the nine in the book is explicitly about assessment. The reason for this lies in the motivation of the Council of Europe, an institution not usually concerned with matters such as language learning.

The aim of the CEFR was to promote language learning in Europe, and to make it easier to compare different levels of language and different qualifications across national borders. This was to be done by providing a set of common guidelines, and the ultimate objective was to encourage mobility and mutual understanding in the whole of Europe.

So how can the CEFR be of practical use to BE trainers wanting to produce professional assessment? It is useful to be familiar with the CEFR itself as well as with several documents which have arisen out of it. These are the *Manual for Relating Language Examinations to the CEFR*, the *European Language Portfolio*, *Dialang* and *English Profile*, a project at the moment under way.

**▶ Further information**

## 2 The CEFR

### How to use the CEFR

The one thing that almost everybody connected with languages in any way knows about the CEFR is that it has six levels, A1 to C2. However what many do not know is not only

that these levels are described in a great deal of detail, but also that there is a wealth of other information in the publication. It is also important to know, and particularly when talking about tests, that the level descriptions are the only ones for these levels. In other words, if B1 is mentioned, then the only valid description of B1 is the one in the CEFR. The levels cannot be defined by individuals or related to sometimes vague ideas of what *intermediate* and *advanced* levels are. It was precisely because terms like *intermediate* and *advanced* are vague and often undefined, that the CEFR was developed. Although adaptations are allowed and even encouraged, these should remain within limits if they are to keep the spirit of the original document. So BE trainers should know what they mean when they refer to the CEFR.

As so many tests, both high- and low-stakes, claim to make statements about the test-taker's ability or skills at a certain CEFR level, these can only be evaluated with knowledge of what the level means. Unfortunately in many cases the level names have come to be used independently of what they really mean. If you want to claim that your test is one of say, oral competence at level B2, then you should equip yourself with a copy of the CEFR (downloadable free from the website of the Council of Europe) and find out what being able to speak at this level means. This is one form of validity. Unfortunately many claims as to CEFR levels are made without much or any knowledge of the document itself. It may have been this state of affairs which led to a further Council of Europe publication – the *Manual for Relating Examinations to the CEFR*.


*The Council of Europe building in Strasbourg, France*

**Did you know?**

It can be assumed that proof of quality assurance will become necessary for all foreign language examinations in the future, if these claim to be related to the CEFR and the competence levels and wish to continue to exist and be recognised in a European context.

So if you want to make a claim for your test or if you have been asked to set a test at a certain level of the CEFR, the first thing to do is to use the CEFR to do this. You will notice that there are boxes throughout the publication like this one. They all start in the same way.

> "Users of the Framework may wish to consider and where appropriate state:
> To what degree which skills are required for the satisfactory accomplishment of the communicative tasks the learners is expected to undertake;
> Which skills can be presupposed and which will need to be developed;
> Which reference aids the learner will need/be equipped/be required to use effectively."
>
> CEFR, p. 93

These show clearly that the CEFR is not intended to be prescriptive but designed to stimulate and encourage thought on certain aspects of language learning.

Next have a look at Chapter 4 of the CEFR (*Language Use and the Language Learner*). Although it is difficult to identify and isolate the parts of the CEFR which are most

relevant to language testing in a BE context, Chapter 4 is certainly one place to start. It defines domains of language use, which could provide the starting point for the broad construct for BE tests. A domain is defined as a sphere of action or area of concern, in other words, where language is used and to what purpose.

The domain for BE is the domain called occupational "in which the person concerned is engaged win his or her job or profession" (CEFR, p. 45). Also in chapter 4 there is a list of questions (CEFR, p. 44) some of which are reprinted here in a slightly modified form:

- What roles will learners have to play in business life?
- What sort of people will they have to deal with?
- What things and topics will they need to refer to and deal with?
- What tasks will they have to carry out?
- Will they have to speak or only read or listen and understand?

All these questions apply to the authenticity of BE tests.

If you do not want to read the entire CEFR from beginning to end and mainly want to know about its usefulness for testing, here are some suggestions for familiarizing yourself with what you need.

- Read the contents page and skim through the book to get a general overview.
- Read Chapter 4.
- Read the global scale on page 24.
- Read the chapter on Assessment.
- Choose one set of descriptive scales (for instance for oral production from page 58 onwards) and read them to get an idea of what a level means in terms of a specific skill.
- Read as many descriptions as possible for the level most relevant to your needs. In the case of BE this will probably be B2 or a higher level.

## What does the CEFR say about assessment?

**Did you know?**

*Assessment* is a broader term than *testing* and includes self- and portfolio assessment as well as continuous assessment. Evaluation is usually used for the evaluation of entire programmes.

The CEFR has a whole chapter (but only one of nine) on assessment. It is explicitly about assessment and not testing or evaluation.

Although the CEFR says little about the practical aspects of formal testing, it still contains much useful information on assessment in general. However the most important thing is that the CEFR itself provides the construct for the validity of a test as soon as you claim that your test is at a certain CEFR level. This means that if you do this, you will have to use the CEFR to start validating your claim.

Each of the levels is described in terms of a large number of aspects of language use by means of can-do statements which say what a language user can do at that level. The 58 descriptive scales describe many different aspects of language. These can be used as a starting point for the development of marking criteria for tests. For instance if you are developing a speaking test, you would probably want to consider fluency and pronunciation.

## Some important points

The CEFR marks a new approach to language learning, one in which communicative competences play a far more important role than accuracy of any sort. This is bound to have a huge effect on testing. Traditional teaching and testing has often been of grammar and detail and has failed to take many of the competences mentioned in the CEFR into account (sociolinguistic appropriateness, flexibility, turntaking, pragmatic competence to name but a few).

The descriptors should be used selectively, especially from level B1 upwards as not all the skills described may be relevant to all learners.

The descriptors in the CEFR are not intended to relate to the way native speakers use language. There are things described at C2 (the highest level in the CEFR) which not all native speakers can do and not all learners will need or want to do. It is important to bear in mind that the descriptors are not intended for the description of or in comparison to native speaker speech.

**The CEFR is not a set of rules**. This is said often enough, but also very often forgotten. It describes language use, and provides descriptions which can be used to design teaching material and tests.

Although the CEFR was originally written in English and many of the examples in it are taken from English, it is not intended for learners of English alone. It refers to language use in general, regardless of the language, and has been translated into several non-European languages.

## What do the levels mean and where does BE start?

Looking more closely at the levels and the level descriptions will help you to decide what you mean by BE, what your language construct is and what level your test is intended for. This is important if you are claiming that your test is at a certain CEFR level. Looking at the level descriptions will help to decide if it is realistic to talk about BE at low levels. Being able to use language to deal with specialist subjects or unfamiliar topics is referred to in the CEFR only at and above level B2. At lower levels than this, language can of course be used in the occupational domain but whether this can be called BE will depend on your definition of what BE is.

## At what level can Business English be tested?

Here is the basic level scheme. There are three main levels, each divided into two. The English names were abandoned, as they were found difficult to translate (and sometimes, as in the case of the word *Vantage*, even difficult to understand) in favour of letters and numbers.

| CEFR levels | | |
|---|---|---|
| Basic user | A1 | Breakthrough |
| | A2 | Waystage |
| Independent user | B1 | Threshold |
| | B2 | Vantage |
| Proficient user | C1 | Proficiency |
| | C2 | Mastery |

In the 58 descriptive scales, in general, there is a descriptor for each level, but in some examples, the top and bottom descriptors are missing because it is difficult to define the skill at low levels and the "highest possible level" in the particular scale is reached before the highest level, C2.

The scales are grouped according to the overall skill (speaking, writing, listening and reading) as well as the strategies needed to be able to do these things. The scales for communicative strategies include individual strategies, like planning, compensating

and monitoring and repair (noticing mistakes and correcting them) as well as interactive strategies such as taking the floor, cooperating and asking for clarification, and intellectual strategies, such as note-taking and processing text. There is a separate set of scales for competences which includes vocabulary range and control, grammatical accuracy, phonological control and orthographic control, and a set of sociolinguistic competences, including sociolinguistic appropriateness, flexibility, turn taking, thematic development and coherence and cohesion. The complete scales can be found on the CD-ROM.

Although the titles of the scales may sometimes seem daunting (e.g. *coherence and cohesion, propositional precision*), the descriptors themselves are written in everyday language and are easy to understand. Some descriptors can be used as the basis for marking criteria for a test. For example, in the scale *propositional precision*, level B2 is described as "Can pass on detailed information reliably" (CEFR, p. 129). This could be used to design criteria for testing whether someone can pass on a telephone message which includes details, for example of a product or an order.

For a test, it is important to identify what the descriptors mean in terms of actual language. The project *English Profile*, at present under way, is intended to provide detailed information about the language which learners can be expected to use at each level.

www. Englishprofile.org

However the CEFR was not developed for the sole purpose of testing, and so, although it can be extremely useful for test setters, it will not provide all the answers to questions test setters may have. The large number of scales makes it necessary to choose the ones for a test. In tests of speaking, there is a limit to what examiners can assess at one time.

▶ Chapter 10

## The main things to remember about the levels

The levels are not based on grammatical progression but on communicative ability and competence, what learners can do rather than what they know. In fact of all the many scales used to describe language, only three are concerned with accuracy and only one with grammar. This is the major difference between the CEFR and a traditional approach, and this can have a dramatic effect on testing, especially if you want to test accuracy, as grammatical accuracy only really plays a role in the CEFR above level B2.

**Example**

Grammatical accuracy B1: "Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influence. Errors occur, but it is clear what he/she is trying to express."

CEFR, p. 114

The scales are not all of equal importance. The descriptors of overall skills like reading and speaking are probably more important than the more specific descriptors of sub-skills. There are some you may choose to ignore, depending on your target group, and some which will be essential. This applies to teaching and of course to testing.

The levels are not equally distant from each other, and learners do not automatically move from one level to the next in each course. This is often a disadvantage, as it is difficult to explain to learners that because they have just done and passed a test at level B1, this does not mean that in a year they can automatically take and pass a test at level B2. If you are designing progress tests, you may wish to assess particular skills at one level in different tests or make it clear to learners that they are making progress even if they have not moved up a whole level since the last test.

The very detailed descriptions of the levels in terms of sub-skills can be extremely useful for testing, if you can set appropriate tasks. It can mean that you can try and test specific skills and competences independently of each other. This is particularly good for diagnostic testing and for setting tests for small groups.

## Where does BE start?

In the overall scales, it is only from B2 upwards that words specifically connected with BE are found – *complex text, abstract topics, technical discussions, detailed text, wide range, topical issue, advantages and disadvantages.* It is fairly easy to see how these relate to BE. Although the word *work* occurs at B1, this is more in the sense of everyday life, and the rest of the global description for B1 has many more expressions associated not with work or business but with the personal domain – *school, leisure, travelling, familiar, personal interest, dreams, hopes, ambitions.*

If it is also assumed that a certain level of accuracy is important for BE, then looking at the three scales for accuracy – grammatical accuracy, phonological control and orthographic control – will show clearly that these only become important at B2 where a relatively high or a good degree of grammatical control is one of the features of the level, as well as a clear natural pronunciation and clear intelligible writing. Looking at these features in relation to the levels above and (particularly) below them will help when setting placement tests to distinguish between what is expected and what the test-takers produce. But remember, these three scales of accuracy are not all there is to the CEFR!

## Which of the skills described are useful or necessary in BE?

Obviously the answer will depend on you and the target group as well as what you want to test, but here are some examples of relevant skills.

- Addressing audiences
- Reports and essays
- Listening as a member of a live audience
- Reading correspondence
- Transactions to obtain goods and services

The detailed descriptions in the CEFR of partial competences provide an extremely useful basis not only for discussions with learners and course planners of learning objectives, but also for test setting, i. e. deciding which of the specific skills are necessary for the learner, should be tested and can be tested.

## How can the descriptors be used to produce tests?

The CEFR levels descriptions can be used to produce BE tests in several different ways. Using the CEFR does not mean that existing tests can no longer be used, but as any test needs to undergo a process of constant revision, this process can take place using the CEFR to see how far what is tested corresponds with its descriptions.

The CEFR can be used to help find the answers to the four basic questions asked in this book about tests. If it is assumed that a BE test is to be set at level B2, testing the achievement of the level in one or more skills at the end of a course:

1 What do I want to test? → Choose the descriptors most relevant to the skill and your learners' needs.
2 What sort of material do I need? → The CEFR provides examples of the type of material for all the skills such as "clear well-structured expositions of complex subjects" (CEFR, p. 62), "clearly structured lecture" (CEFR, p. 96), as well as details

of what is meant by these. These sometimes rather abstract descriptions need to be related to actual text and material available.

3 How can I make sure my test is reliable? → If you want to attempt to align your test to the CEFR, then keep comparing the test, the performance of the test-takers and the results you issue as often as you can.

## Does the CEFR need adaptation to BE and is this allowed?

The answers to these two questions are "only partly" and "yes". The reasons for this are firstly that the CEFR was produced to cover as many different settings as possible. These have been defined broadly as domains, but clearly there are many other more specific areas which the CEFR could be used in. Many of the expressions used are general and would need to be "filled out" for specific contexts. Business is one such context, even though it is clearly a part of the occupational domain. The CEFR says explicitly:

> "Users of the Framework may wish to consider and where appropriate state:
> - in which domains the learner will need/be equipped/be required to operate;
> - the situations which the learner will need/be equipped/be required to handle;
> - the locations/institutions/organisations, persons, objects, events and actions with which the learner will be concerned."
>
> CEFR, p. 47

The CEFR includes a set of descriptive categories for each domain where lists of locations, institutions, persons, objects, events, operations and texts are given. Many of those listed for the occupational domain will be closely related to a definition of BE.

Although obviously you are not free to redefine the levels, there are several places where the CEFR may not be specific enough for some needs.

Just to take two examples of descriptors from level B2:

> Information exchange B2:
> "Can understand and exchange complex information and advice on the full range of matters related to his/her occupational role."
>
> CEFR, p. 81

Obviously here the range of matters related to *his/her occupational role* needs defining more closely, depending on whether they are related to a train attendant, a tax advisor, a hospital nurse or a teacher, and the CEFR has in fact been adapted to some of these roles. One example is the publication *Working in Europe*, produced by the German Chamber of Industry and Trade (DIHT), now available only in German.

> Thematic Development B2:
> "Can develop a clear description or narrative, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion."
>
> CEFR, p. 125

The questions to be asked here when designing a test would be what a clear description or narrative in a BE context is. It could mean, for example, a description of the history of a company to be given to a group of visitors.

It is exactly these definitions and details which add to the specific context of a test and increase its validity for that context.

## Portfolio

The *European Language Portfolio* and the CEFR were developed at the same time and both have the reference levels as their basis. There are a large number of portfolios which have been recognised by the Council of Europe. These have been developed for different target groups, ranging from primary school pupils to adults in particular countries. The main aim of the portfolio is to develop learners' autonomy. Another aim is to provide a format for recording learning experiences, knowledge, skills and qualifications, at all levels. Although there are several connections between the portfolio and assessment, its relevance to BE will have to be decided individually by trainers.

www.coe.int/portfolio

## Manual

Although the CEFR descriptors can be used themselves to relate learning content, class assessment and examinations to the CEFR levels, there have been several further developments making this easier and more accessible to teachers and trainers. One of these is the *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages (CEFR) and Reference Supplement* designed "to help the providers of examinations to develop, apply and report transparent, practical procedures in a cumulative process of continuing improvement in order to situate their examination(s) in relation to the CEFR. The *Manual* aims to:

- "contribute to competence building in the area of linking assessments to the CEFR;
- encourage increased transparency on the part of examination providers;
- encourage the development of both formal and informal national and international networks of institutions and experts".

from www.coe.int where the complete Manual can also be downloaded.

The *Manual* provides a structure for test setters for designing specifications relating to the CEFR, for formalising the way in which their exam relates to the CEFR levels and for standardising judgements on test-takers' performance in tests of speaking and writing. Video examples of speaking levels and samples of test-takers' writing, both calibrated at the levels of the CEFR, are also available from the Council of Europe website. There is also information in the *Manual* on providing empirical validation. Although this is something that is of greater interest to external examination providers, BE trainers and heads of language departments at universities may also have to provide some evidence of how their tests relate to the CEFR and how valid and reliable they are. Fairly simple methods of doing this are described in the *Manual*.

The *Manual* probably contains more detail than most BE trainers need, but it can be considered essential reading for anyone really wanting to link tests and exams seriously to the CEFR. For standardised tests in institutions such as business schools or universities, it could be used in a project group and the information disseminated in training sessions, as suggested in the publication.

## Dialang

One further development which deserves mention is *Dialang*, a diagnostic test available on the Internet in 14 languages. It is intended as just that, for learner's self-assessment, and gives the test-taker information on strong and weak points and advice for further

learning. It should not be used for any other sort of testing, but learners can be guided to it for an initial assessment of their level before proceeding to a further test. *Dialang* was developed by testing experts who had been closely involved with the CEFR, with financial support from the European Union and made freely available to anyone wanting to use it. There are tests for reading, writing, listening, grammatical structures and vocabulary and detailed explanations for doing these in all 14 languages.

One important thing to realise is that *Dialang* is not a certificate and cannot be used to certify a level of language or language proficiency. *Dialang* gives language learners information about their strengths and weaknesses in a language, however they have learnt the language. It is available to anyone, but the content of the items may be more suitable for teenagers and adults. *Dialang* uses self-assessment, feedback and diagnosis to provide learners with a statement about their language skills and needs. A test may take about 30 to 45 minutes to do, but there is no time limit.

www.lancs.ac.uk/researchenterprise/dialang/about

■ The CEFR had the very ambitious aim of providing a transparent and coherent framework which could be used by all language professionals, teachers, trainers and test setters as well as being accessible to stake-holders such as parents and employers. Whether and to what extent you make use of all the information and guidelines in the CEFR itself, the portfolio, the *Manual* or encourage your learners to use *Dialang*, the CEFR is so ubiquitous nowadays that almost anything done in languages claims to be at a certain CEFR level. If you intend to make this claim for your test, then the more you know about the levels, the more reliable and valid your test is likely to be. The important thing is not to regard giving a test a name like B2 or C1 simply as a matter of labelling, but of taking it seriously and making sure that any claims made are backed up with the relevant information. ■

### To think about:

- Think of any language you have learnt and look at some of the can-do statements to decide what level you are at.
- Which of the descriptive scales in the CEFR are most relevant to your learners?
- How important is grammatical accuracy to you in tests of speaking and writing?

CD-ROM

*"Everyone hates them, yet at the same time, we all recognize their importance for measuring our achievement."*

Alan Maley in May, Peter (1996): Exam Classes. OUP, Foreword

# 6 Choosing an exam

## SEW, LCCI, TELC, BEC, PTE, TOEFL … which is best?

- Why do your learners want to take an external exam?
- Which exam is the best for your learners?
- Which exams are available?

Most BE trainers are asked at some time or another to choose or recommend an exam for learners to take in order for them to gain some sort of certification of their level in English. The question is usually formulated "Which is the best exam to take?" or "Which exam is recognised most widely?". Clearly no one can legitimately be expected to know the details of all the BE exams available on the market, and often the learner's request is vague. The more specific the learners' or the institution's request is, the easier your task will be to choose and/or recommend the best exam for the specific needs. This chapter will help you when you are placed in such a position and will give you some of the criteria you need to choose or to help learners to choose an exam. You may also be asked to recommend an exam for your institution in which case similar considerations will apply.

## 1 Why does the learner want to take an external exam?

The first thing to do is to find out exactly why an external exam or qualification is necessary or desirable. The more you know about this the easier the first step will be.

One reason may be that a learner is applying for something (a new job, a course, a study programme) and wants to or has to provide some proof of his/her knowledge and skills

in English. The first thing to find out is if it is absolutely necessary for the learner to take an external exam. While not wanting to discourage learners from taking exams, a certain amount of effort and expense will be necessary in order for them to do this and you may discover that this is not worthwhile. A school qualification, proof that the learner has attended English courses or a recommendation from the trainer may be enough. Possibly there is also a test connected with the application (for example, part of the job interview is in English or there is an admissions procedure which involves an English test). In cases like this, it may make more sense to spend time preparing and practising for this rather than for an external exam.

## 2 Questions

### Which exam is recognised?

The first thing that must be said is that there is no general "recognition" of exams. Any recognition will depend on the purposes the qualification is needed for and the institutions involved. The fact that an exam claims to be at a level of the CEFR does not mean that it is in any way recognised by the Council of Europe. This is often insinuated by words like "at the recognised CEFR level B1" or "recognised as B2", but this does not mean official recognition by the Council of Europe (or the European Union!). This recognition of exams may come about in the future, and there is discussion of this going on, but at present it does not exist. Usually recognition is specific: an employer may recognise a particular certificate as proof of English competence at a certain level, or a university may recognise the results of a particular exam for admission to a study programme. This is what is meant by the claims of exam boards that their exams are recognised "in over 100 countries" or "by universities worldwide". There are of course bodies which recognise all sorts of things, including exams, sometimes generally, sometimes for certain purposes. Some of these bodies are prestigious, some less so. If an exam or a test is "recognised" by someone or something, it is worth finding out about the recognising body. An exam is usually considered as widely or well recognised if a large number of institutions accept its results as proof of competence in English. This probably just means that it is well-known.

If an exam is well-known, then the advantages are clear: the learner will be able to use the results for a number of different purposes and in a number of different situations. However, the more widely recognised, i.e. well-known, the exam is and the larger the target group it is intended for, the smaller the chances are that its content will suit the exact needs of any single learner or match any single learner's specific skills and competences. This means that practising for the exam will be necessary. While this need not necessarily be a waste of time, this time may be more profitably spent on other practice for the new job or course of studies. All this can be established by first finding out whether the learner really needs to take an external exam and get an external qualification.

▶ Chapter 16

**Remember**

The criteria for a good test apply just as much (if not more) to external exams.

This chapter assumes that learners want to take an exam (for whatever reason) and can cope with the effort involved in this undertaking. You need neither dissuade them from doing this nor persuade them to do it, but should find out what their reasons are for taking an external exam so that you can help them in the best possible way.

This applies as well if you are comparing two or more exams or choosing an exam from the range available. The procedure is similar in all these cases. What you need to know

in order to best help your learners is much the same whether you are only looking at one exam or making a choice from several.

The following questions apply when you are choosing an exam to recommend. This is the most extensive task you can be faced with.

▶ Chapter 1

### When do learners want to take the exam?

You will need to know whether they want to take an external exam fairly soon in order to have proof of their level of English now, or whether they want to continue learning and take an exam, say in a year or so (or at the end of this course or after another course). This way, you can decide on the answer to the next question.

### What level should the exam be at?

It may be better for learners to take an exam at a low level and may get a high grade, or at a high level and get a higher-level qualification which will probably have more value in the long-term. This may depend on whether they want to aim at the highest grade possible or are happy with a lower grade or just a pass.

You may also be reading this book because you are teaching an exam course, at the end of which learners will be taking a particular exam. Even though there may be people in the course who have no intention of taking the exam at the end, you should still know as much as possible about the particular exam in order to be able to give those who do want to take it the best possible preparation. In this case, your research will be limited to the one specific exam. You can probably answer some of the questions below yourself, but some answers will have to be given by the learners.

Your recommendation of which exam they should take and at which level will depend on their wishes and your assessment of their abilities. Assuming you have an idea of their level, the first thing to do will be to look at exams at this level. This will narrow your choice down a bit. BE courses do not usually start below level B1 which is reflected in the range of BE exams available. This will involve looking at the description of the target group the exam is aimed at and the sample material available, to see whether your learners are already at the level of the exam and/or how much more they will have to do in order to get there. Both these things should be described on the exam board's website.

### How much time do you have to prepare learners for the exam?

It is unlikely that learners will want to take an exam immediately, but you will need to know how much time you need to spend on specific exam preparation. Remember that a good exam will not be testing whether, for instance, learners can answer multiple-choice or other types of questions, but rather whether they have certain knowledge, skills and competences in BE. Although a certain amount of specific exam preparation may be necessary, this should not be the only thing you think about. The less specific exam preparation you have to do, the better. The time you have available will also depend on the learners' needs and the availability of the exam.

**Did you know?**

Although several exam boards allow retakes as many times as the test-taker wants, there are relatively few people who retake exams simply to get a better grade if they don't have to.

You will have to find out whether an exam can be taken on demand or at fairly short notice if you have a group of learners, or whether an individual learner can join a group already taking the exam. This information will be relevant even if learners are planning to take an exam at a later date and more planning can take place. Different exam boards have different arrangements for taking their exams. Some offer exams on only two or three dates a year at official centres. Others are more flexible, offering several dates a year or dates on demand, while yet other exams can be taken fairly regularly. External exams

are usually taken at official centres, so you (or the learners) will have to find out where these are. Although availability should not be your main reason for choosing an exam, it may well influence the choice finally made for the learner. There is no point in finding the perfect exam for a learner if you have just missed the exam date and the next one is in a year in central Africa! This is of course unlikely, but may demonstrate the point.

## How much do learners want to pay?

This is another question which should not ultimately influence the choice of the best exam. However it may play an important role, depending on the learners' reasons for wanting to take the exam in the first place. Details of prices usually depend on the locality and the institution at which the exam can be taken as well as sometimes special conditions. They are not usually provided on the exam boards' websites or only with fairly vague indications and guidelines.

### Which exam is best?

So far the questions we have asked require little specialist knowledge on your part. They are not however unimportant, as they may influence the final choice of exam. What learners will want to know from you is, however, most likely which exam is "best" for them. Apart from the question of recognition, which we have looked at above, there are several other factors you should think about.

### What material will I need?

It is clear from the above that one of the most important things an exam board should provide is sample exam material, probably in the form of practice papers. Sometimes a number of past papers are published on the website. Sample exam material is the absolute minimum you need, but this on its own will not be enough to find out what the exam is really testing. You will also need a *syllabus*, so you know the entire range from which the exam tasks are taken. The syllabus may include a word list or a list of functions and grammatical structures which may be tested in the exam. Remember that a test is usually a selection from a larger amount of material. The syllabus may be called something different but will basically be a list of language – words, functions, grammar etc. It is a part of the specifications.

▶ Chapter 2

**Remember**

Find out what your learners want to know. Some exam boards publish lists of FAQs where you can find the answers.

An exam board may not publish its entire specifications as not all parts are of equal interest to everyone, but you should find enough material to answer your and your learners' questions.

Of all this, probably the most immediately useful, and therefore essential, information will be in past papers and sample material. Any reputable exam board really should provide this information so you and the learners can see what the exam looks like. In fact, a good exam board should be able to answer all your questions, without revealing the actual questions to be asked in the exam.

### Procedure and criteria

As well as the material for the receptive tasks in the exam, you will also need information on what happens in the oral exam and how the productive (written and oral) tasks are marked. It can be very helpful to attend an examiner training session if these are provided by the exam board. Examiners and prospective examiners know best of all what happens in the oral exam and how it is marked. There may be video material available for this as well as examples of written performance with grades and marks for writing tasks. When you are marking your learners' practice exams, try and stick to these and not to interpret them yourself.

If you want to actually work as an examiner, bear in mind that you will probably not be able to examine your own learners. Most exam boards do not allow this.

### Results

Learners will almost certainly want to know about the results. Their main questions will probably be:

- When will I get the results?
- Will I get a grade or can I just pass or fail?
- How are the grades calculated?
- Do I just get a certificate or can I have a breakdown of my results?
- Can I appeal against my results if I think they are wrong or not high enough?
- Can I see my paper and discuss the marks?
- When can I retake the exam?
- Can I retake one part if I only fail that part?
- How long are the results valid for?

All this information should be available. It is sometimes contained in the examination regulations, which should also be published. The results should not take too long to reach the test-taker, although probably the larger the exam board, the longer the results will take. Some exam boards promise results within a certain time. The test-takers should get a breakdown of their results, whether they have passed or failed, and a certificate if they pass. Test-takers should have the opportunity to retake the exam if they fail and may be able to retake it at a later date to get a better grade. They may also be able to appeal and to see their papers and discuss the marking.

However, not all exam boards allow retakes simply to get a better grade. Some exam boards do not have an appeals procedure and some charge for appeals. There are often organisational and even geographical reasons for not allowing test-takers to see their papers and discuss their results. It is worth finding out all this beforehand. Usually learners are happier if they have all the information, even if it is not exactly what they want. They then feel that they are being treated fairly.

**To think about**

Fairness is an important feature of good exams and tests. One way of being fair is to provide all the information necessary for someone taking the test.

The results of some exams become invalid after a certain time. The most notable of these exams is TOEFL (not described here as it is not a BE exam). This is explained by attrition, in this case, the process of losing language ability and skills if they are not used. It is therefore the subject of a question often asked by learners, and you can expect to get an answer to it from the exam board.

▶ Chapter 15

Even if the exam result does not become invalid, it is in fact a snapshot of the learner's performance on a particular day so the results cannot be considered as reliable for ever, whether the learner's English becomes "better" or "worse".

### Where do I get the information I need?

The most obvious answer of course is the Internet, although many exam boards also publish a large amount of print material. This can be sent by post or is available at conferences and workshops. Exam boards often run sessions at conferences or on individual pre-arranged dates. Large exam boards usually have local representatives who can give you a great deal of information. The information is sometimes presented as information for test-takers, trainers and institutions but if you are choosing an exam, you should look at all the material. Learners of course will probably only need the information relevant to them. Bear in mind that it may be easier to go through print material with your learners. It may also help if the information is available in the learners' language.

As most exam boards want to sell their exams, they will of course present them in a favourable light. However there are a few things which may be worth looking out for on websites, especially those of less well-known exams:

- Words such as *perfect* and *ideal*. These promise too much. No exam is perfect or ideal for everyone.
- Poor English in texts on the website. This is probably only in the case of exam boards not based in an English-speaking country. However it should raise some doubts in your mind as to the quality of the language in the exams.
- Material, such as sample papers, which is only available at a price or not at all. Although exam boards are often commercial enterprises, there is a certain standard and quality of information which should be easily accessible and available at no cost to those who need it.

## 3 Overview of Business English language tests

The following information on some of the most widely-taken international exams which are relevant for BE trainers may help you to make an initial choice. As well as this, there are several exams and tests offered by smaller institutions (usually language schools or local organisations) which do not, however, figure prominently as recognised qualifications and so have not been included here. However, the criteria in this chapter can be applied just as well to these.

All these exam boards offer a range of English exams for both general and Business English. All of them provide information (to a varying extent) on the following:

- Links to the CEFR levels
- Exam content
- Exam preparation
- Research reports
- Teaching material
- Recognition
- Administrative procedure, centres etc.

Exam boards which do not provide any of this information have not been included and only BE exams are mentioned here. There are of course other exams, for instance recognised as university entrance qualifications (CPE, TOEFL, IELTS and PTE Academic are the best known at present), but these are not specifically tests of BE. They are however definitely worth considering as alternatives, depending on your learners' needs.

*Which is the best test to choose?*

The information given here is taken from information provided by the exam boards and is neither comprehensive nor exclusive as all the exam boards mentioned provide a large amount of information, often as downloads on their websites.

### Cambridge ESOL

The Cambridge ESOL (English for Speakers of Other Languages) examinations are probably the most widely-known, and information is freely available. There are exam

centres all over the world. Cambridge ESOL provides information on all the important aspects of exams and publishes extensive research reports. BE trainers preparing test-takers for Cambridge ESOL exams will find all the information they need in the handbooks published for each exam. These are available as downloads or in print form. Information is also available in different languages. Cambridge ESOL provides a list of companies which recognise their BE exams.

The BEC (Business English Certificates) exams are at three levels from B1 to C1. All four skills are tested and the written parts can be taken as a computer-based as well as paper-based test. The texts are taken from a variety of sources from English-speaking countries. Sample papers are provided, as well as examination reports and video material to familiarise test-takers with the format (these are not free of charge). There are a number of course books to prepare test-takers for BEC exams. Cambridge ESOL does not advise on or endorse these but does provide online support for teachers preparing test-takers for the exams. www.CambridgeESOL.org/BEC

**ICFE**

ICFE (The International Certificate in Financial English) is a test for finance professionals working internationally. It does not test accounting but rather the language skills people working or intending to work in finance need. It was developed in association with the Association of Chartered Certified Accountants and is set at levels B2 and C1 of the CEFR. There is also some teaching material available to prepare test-takers for this exam. www.financialEnglish.org

**ILEC**

ILEC (The International Legal English Certificate) contains tasks taken from the work of legal professionals. It is not a test of law, but tests the language skills needed by lawyers and other members of the legal profession or those working in a legal context for communication in English. It is produced in cooperation with the TransLegal group and is set at levels B2 and C1 of the CEFR. Teaching material is available. www.legalEnglishTest.org

**BULATS**

BULATS (Business Language Testing Service) is not an external examination like the others mentioned here, but is an English language evaluation tool for organisations and companies. It assesses all four skills and is linked to the CEFR and other Cambridge ESOL exams. It is offered as the BULATS Computer Test, the BULATS Standard Test, the BULATS Writing Test and the BULATS Speaking Test and is available in several languages. BULATS tests language skills in a range of occupational situations, such as those technicians, secretaries or managers might find themselves in. It can also be useful for university students. Information on the test and how learners can prepare for it is available in several languages. www.BULATS.org

**IELTS**

IELTS (The International Language Testing System) is also managed by Cambridge ESOL, and although not strictly a BE exam in the same sense as the others mentioned here, is also worth considering. IELTS is recognised as a university entrance qualification in many countries and is also recognised by professional bodies and immigration authorities. IELTS is not recommended for test-takers under 16. The results are provided in the form of points, and the number of points necessary varies according to the recognising body. The range of CEFR levels is from B1 to C2 www.ielts.org

**LCCI**

The LCCI (London Chamber of Commerce and Industry) international qualifications are offered by EDI, formed by a merger of LCC and Goal Assessments in 2003. English is

one of the subjects in a range of professional subjects offered. LCCI also offer exams leading to professional qualifications (non-language) in finance and business administration. Each exam has a product guide and a full syllabus. Other resources are also available such as course material published by LCCI itself and officially recognised as exam preparation material. LCCI exams are recognised by several government ministries and institutions worldwide.

www.lcci.org.uk

**EFB**

EFB (English for Business) is offered at all levels from A1 to C2, one exam generally covering two levels. The reading and writing sections of the tests are compulsory and the listening and speaking tasks are optional at all levels.

**SEFIC**

SEFIC (Spoken English for Industry and Commerce) is an exam offered at levels A2 to C1. As the name suggests, this is only an oral exam which consists of a one-to-one interview with an examiner. The length of the interview and the complexity of the tasks increase with the level. English for Tourism is also offered by LCCI at levels A2 and B1. The two components, written and spoken English, are available separately or together for each level.

**City and Guilds**

City and Guilds is a major provider of vocational qualifications in the UK and also has a large number of centres worldwide. The exams are available on demand at several centres.

www.cityandguilds.com

**EBC**

EBC (English for Business Communications) exams are exams with only written tasks at three levels, A2, B2 and C2. The exam tasks involve reading documents and responding to them appropriately. These awards, at three levels, equivalent to A2, B2 and C2 test comprehension and writing of business texts.

**SETB**

SETB (Spoken English Test for Business) is for those learners who specifically want a test of spoken BE. It is available at the same three levels as EBC.

**English for Office Skills**

These examinations are also intended for test-takers whose first language is English. They test the ability to use correct and accurate English in office situations and include proof-reading and spelling tasks.

**Trinity College London**

Trinity College London is an international examinations board and provides qualifications in the English language, music and the performing and creative arts. Exams can be taken worldwide at Trinity centres. In 2006, some Trinity exams were linked to the CEFR levels by means of extensive research with an independent consultant. The report is available on the Trinity website. Details of recognition in several countries are provided.

www.trinitycollege.co.uk

**SEW**

SEW (Spoken English for Work) is a range of exams in BE open to anyone over 16. The exams are set at four levels (B1 to C1) and test spoken skills in a number of different business situations. A syllabus is provided for teachers' use and a DVD is available with examples of test-taker performance.

**telc**

The telc exams were developed at Volkshochschulen in Germany and can be taken at these. The abbreviation telc is now used as a brand name but originally stood for The European Language Certificates which were organised and run by WBT Weiterbildungs-Testsysteme GmbH. telc exams can be taken in several European countries and in China. They are available in different languages and at all levels of the CEFR. BE exams are available at two levels, B1 and B2. English Hotel and Restaurant (B1) and English Technical (B2) are also available. The exams test all four skills, and sample material is available as well as descriptions of learning objectives and the test format. www.telc.net

## 4 Some comparisons

Cambridge ESOL BEC exams are probably most suitable for less specific needs and most useful for learners looking for a general BE qualification which is well-known and recognised by a large number of institutions. For learners working in law or accounting and finance, ILEC and ICFE may be more suitable, provided the learners are at a level high enough. As well as being useful for BE learners, telc exams may also be suitable for those in the hotel industry or in technical occupations. Trinity SEW is an exam which only tests speaking so is suitable for learners wanting to certify only their speaking skills whereas City and Guilds EBC only tests writing. BULATS and IELTS will probably only be taken under certain circumstances, i.e. if the particular qualification is required.

If your view of language is a communicative one and you have adopted the CEFR approach, it may also be worth asking the following questions about external exams and tests:

- Is language used in international encounters and intercultural communication tested rather than in only country-specific or Anglo-Saxon contexts?
- Are relevant skills evaluated and not only or largely knowledge of grammar and pronunciation?
- Do the test-takers adopt realistic roles in the tasks, whether these are in tests of reading, writing or speaking?
- Are the texts authentic and the roles allotted and tasks set communicative?

Depending on the nationality or cultural background of your learners, it may also be useful to consider how far the texts and topics will appeal to them and how much they will be able to relate to these. Topics like holidays, weekends, friendship with animals, concepts of solitude and peace may be related more closely in the exam to a western way of life than an Asian one and therefore be more difficult for test-takers of different cultural backgrounds to deal with. Questions like "Are you happy to do this?" may confuse test-takers from non-Western backgrounds.

All this information should be accessible in the sample papers provided.

## 5 Some shortcomings

Let us now look at a few possible shortcomings external exams may have as far as you and your learners are concerned. These may make working towards them more time-consuming or may even mean they are unsuitable. If you are choosing an exam, they may be reasons for not choosing a particular one.

▶ Chapter 16

**There may not be enough information available:** If this is the case it may be an idea to write to the exam board and ask for what you want. Even though a CEFR level may be mentioned, you should not rely on this too much. Each level in the CEFR represents in any case a range of ability, and you will need to judge at what point in the particular level the exam tasks seem to be. Sample material is essential and represents the absolute minimum an exam board should provide.

**You find mistakes in the test material:** This may be a reason for not choosing the exam, but if the exam is otherwise suitable, it may be worth contacting the exam board. It is only through feedback of this kind that exams can be improved.

**The tasks seem irrelevant:** Think about why they might have been included, bearing in mind the criteria for good tests. They may be in an exam for reasons of reliability and practicability. Big exam boards do research to set the best possible tests, and often the methods and results of the research are published. Reading this may give you more insight into the rationale behind the tasks. This rationale may also be explained in published specifications.

**The time for the tasks seems too short or too long for your learners:** Think about whether your learners are really at the right level. The time for completing an exam task is usually set for good reasons, among them the time an effective learner at that level should need, for instance to read a text and answer questions on it. It may be worth considering an exam at a higher or lower level.

▶ Chapter 16

These shortcomings are not necessarily reasons for not choosing a particular exam, but may mean that a certain amount of adjustment in your teaching is necessary.

**Remember**

Remember that you will not be marking the papers, so it is neither your or your learners' subjective judgement which will count, but rather the criteria against which the paper is marked. In most cases the rater's decision is final!

One thing to remember is that external exams will be different in many respects from tests you set yourself. They may look similar but are, if they are good exams, based on extensive research as well as complex and elaborate validation procedures. As they are designed for large numbers of test-takers, not all parts of any one exam will be entirely suitable for your learners. You may have to adjust your teaching so that learners are trained to do a part of the exam they do not need for their work, but have to pass in order to get the qualification. Learners should be made aware of test-taking techniques, and you should find out about the criteria for marking written and oral parts so that you can advise your learners best on what final result they are likely to get in the exam.

■ Whatever the differences and similarities between external exams offered by international examining bodies, placement tests made available by publishers and classroom tests designed by colleagues, the process of choosing an exam from the range of BE exams on offer can be interesting as it will show you the different views of BE that exist. Looking at the details of different external tests and exams and the background to them will also be very helpful in designing and improving your own tests as well as those of colleagues. There is always much more to a test than simply the questions and tasks given to the test-takers! ■

**CD-ROM**

## To think about:

- What do you need to know about the exam in question?
- What do your learners need to do in English?
- Which parts of the exam will be most helpful to them?

*"… to completely analyze what we do when we read would be almost the acme of a psychologist's dream for it would be to describe very many of the most intricate workings of the human mind …"*

Huey, E.B. (1968, originally published in 1908): The Psychology and Pedagogy of Reading. CUP, p. 4

# 7 Reading

## Characteristics of reading tests

- What makes a test of reading easy or more difficult?
- How should texts be chosen?
- What reading skills are being tested?
- What are the best sorts of questions to ask?

A reading comprehension test usually consists of a text or texts and questions. Clearly the process of setting the test is different from setting a test of grammar or vocabulary as first a text has to be found which is suitable and then questions set to test comprehension of the text. This chapter will deal with these two issues as well as other specific features of tests of reading.

### 1 What makes a test of reading easier or more difficult?

Here are some things to consider about the reader, which can make a test of reading easy or difficult for a particular person:

- Someone who is an efficient reader in his/her own language will probably be able to apply the techniques of efficient reading to English. Efficient readers usually read a lot. It may be the case that some learners do not read much and so have not been able to become efficient readers.
- Knowledge of the topic of the text and familiarity with the text type will make a text easier to understand. This is easier to deal with by choosing texts on topics which

learners are familiar with and are used to reading. This affects not only the topic but also the type of text, its source and layout.

- Knowing all the words in a text will make it easier to understand. In fact, if more than 10 % of the words in a text are completely unknown to the reader, the text will be more or less impossible to understand.
  It is possible to guess up to this amount from the context, but it should not be forgotten that deduction of meaning from context is a skill which not all test-takers may have and expecting this in a test can be unfair. Unless you are sure that the words which you want test-takers to guess are really unknown to all of the test-takers, you will be giving those who happen to know the words an unfair advantage. It is best to assume or require that test-takers will know all the words in the text. If you are not sure, you can explain or translate individual words in a footnote.

**Did you know?**

One method of measuring text difficulty is the Flesch-Kincaid index which uses length of sentence, number of syllables in a words to calculate at what age someone could read and understand a text. Like all statistical methods it has its limitations.

The following will contribute to the difficulty of a text:

- Vocabulary
- Sentence structure
- Sentence length
- Sentence complexity
- Text coherence
- Text structure
- Familiarity with topic
- Appropriateness for reader
- Quality of writing

Short sentences with a simple structure and familiar words are considered easier to read than long complicated sentences. Texts which address the reader are usually perceived as easier to read, as are texts which have visual support or "white space" in the layout. A reason for reading in the form of a question or headline also makes a text seem easier. A chronological sequence of events also makes understanding a text easier. Well-written texts are usually easier to understand, in fact ease of comprehension is a feature of good texts.

All these things will have to be considered, whether it is the test setters' intention to make the text easy or difficult.

The difficulty of a reading test depends not only on the text but also on the task or questions.

**Example**

For instance, the simplest test of reading in a foreign language is simply to recognize the language (Bulgarian or Russian? Chinese or Japanese?).
A second very simple task is identifying proper names (people, towns, countries) in a text.

Reading can only be tested at one remove. We cannot test the cognitive process of reading but can only see whether the answers given to test questions are correct. This is why it is essential to make sure that as few other skills are involved as possible, such as general knowledge or common sense. One way of checking this is by asking someone to try and answer the questions without reading the texts. If it is possible to get the correct answers in this way, then the test should be discarded or edited.

**Remember**

If test-takers have to formulate a response in speaking or writing, they must know how much to say or write, and this must be made clear in the instructions.

If open questions are used, for instance if test-takers have to write answers to comprehension questions, spelling and/or grammar mistakes will have to be accounted for in some way. Will they affect the final score? This will mean writing appropriate criteria for these.

There is a current trend towards testing reading in combination with the enabling skills, which include grammar and vocabulary, as there is a high correlation between these and reading skills. However in terms of pragmatic reading skills of the type required in BE, there is a good case for testing reading as a separate skill. This will also depend on the type of test, diagnostic or achievement, for instance.

## 2 How do you start?

When we test reading, we want to find out if test-takers have understood a text and we want them to prove this. Clearly it is not enough in a test to ask "Have you understood the text?" How much proof do we need and how much can we expect?

Identify the type of text you need, the skills involved and the questions which will best suit these two things.

### What sort of texts do you need?

> "Successful choice of texts depends ultimately on experience, judgement, and a certain amount of common sense."
>
> Hughes, A.: Stages of Test Development.
> In: Hughes, A. (2003): Testing for Language Teachers. CUP, p. 58–74

The best sorts of texts to choose for tests of reading are those which your learners will have to deal with in their business life.

What are some of the typical features of these texts? You will need examples of the texts to answer this. This means such things as length of the text as a whole, of sentences, of paragraphs as well as the structure of the text, headings, captions, layout (bullet points, tables) and non-text features, such as pictures and fonts.

Where can you find the texts? Are they generally available? Will you need to consider copyright? Can you write texts yourself?

What is the register used in the texts? For instance, is it formal or informal, journalistic or does it reflect corporate culture?

What language items (vocabulary, grammar) are typically used?

Which other skills do learners need (business, mathematical, general knowledge, cultural)?

On the face of it, there may seem to be no or little difference between the texts you choose for use in class and those used in tests. However looking at exam texts more closely will show that they are in fact quite different from texts used in class.

Some of the main criticisms of exam texts are:
- They are boring, i. e. uninteresting to the reader.
- They are not topical or up-to-date.
- They are inauthentic, the language is sanitised and there is little unfamiliar language.

These criticisms are often unjustified but it is easy to see how they arise. The important thing to remember is that a test is not the same as a classroom activity. The function of the test is not to have fun nor is it to learn something, but to test the learner's ability in a meaningful way. Obviously the test setter is interested in finding texts which are easy to set tasks on and mark objectively, but practicability should not be the only reason for the choice of text. However texts which fulfil the criteria of validity, reliability and are also authentic are not always easy to find and so can (but need not necessarily) lay themselves open to the criticisms made above.

Why should this be so? If exam texts are all these things, this seems to contradict what has been said about authenticity and interaction. It is in fact precisely this which makes it important to select texts for tests of reading carefully.

Interesting texts which provoke discussion are usually very useful in class but what can they be used to test? A good test of understanding might be to let test-takers take part in a discussion, but how will things like deviation from the text be dealt with in marking? How will you grade a misunderstanding of one part of the text, perhaps unimportant for a discussion of the issue and which can be corrected during a discussion? How will you separate the grades for understanding from those for production?

Texts which provide or assume cultural background knowledge may be useful if this is what we are teaching or want to talk about in class, but would a test of reading be the best way of testing this knowledge? Do we want to test knowledge or reading skills?

**Remember**

Although a test is rarely fun, it can be a useful and even enjoyable experience for a test-taker. This is something we should bear in mind. Tests should certainly not trick candidates or be unfair.

There is no reason as such why topical texts should not be used in a test, except that the test-takers' knowledge of the issue will always play a role in answering the questions. Either they will know nothing about the topic and may feel at a disadvantage (and how you feel in a test is important) or their knowledge will interfere with what is said in the text in some way (helpful or not). A practical reason for not using texts which are highly topical is that they can quickly become out of date. This inevitably makes them less interesting and less able to engage the reader. As the entire test-setting process takes some time, this may even happen before the test takes place! If not, it almost certainly means that the text cannot be used again.

However the main problem with choosing texts for tests of reading is that the possibly crucial role played by knowledge of the subject must be taken into account. This is precisely what leads to the choice of bland texts, often on invented topics or situations and also often based on personal or individual stories, in many publicly available exams. There is also the question of bias and fairness, e. g. texts about cosmetics are more likely to appeal to women. The test-taker cannot choose between topics and cannot react to an unsuitable text as he/she could in real life. Even in some tests, it is possible to react to an unsuitable topic without endangering one's results, i. e. in an oral exam or test of writing, it may be possible to choose a topic. However the more you know about your target group, the easier it will be avoid these pitfalls and the better your test will be.

▶ Chapter 15

When you choose a text, make sure it is a valid text for the test. Think carefully about how close it is to what test-takers have to read in real life. Then ask colleagues or friends what they think of it and, if possible, try it out on learners beforehand. Talk to as many people about it as possible. These are the essential steps to test-setting and for tests of reading; they apply both to choosing the text and setting the questions.

Here are some of the differences between texts for testing and texts for teaching.

| Texts for teaching | Why? | Why not for testing? | Texts for testing |
|---|---|---|---|
| • May be about controversial topics | • Can lead to a lively discussion in class | • May offend or upset some test-takers and make it difficult for them to do the test<br>• May provoke feelings of agreement or disagreement which can make it difficult to concentrate on the task | • Neutral topics on which understanding does not depend on personal feelings |
| • May assume cultural knowledge | • Can be explained in class and any gaps filled | • Answers may depend on areas test-takers are not familiar with | • Chosen to appeal to all cultural and age groups |
| • May have many unfamiliar words | • Words can be explained or looked up in class | • Can be used if the test is of deductive skills. However it must then be made certain that no test-takers already know the words or have additional knowledge (e. g. of another language) which might help them | • It is assumed that test-takers know all the language in the test which they need to pass the test. The test only tests reading skills, which makes it more valid |

The ideal situation for the test setter is one with low numbers of test-takers, a homogenous easily-definable group and large resources available for setting and marking. All test-takers will know the same amount about the subject and setting of the texts. Criteria can be developed for the marking of open-ended items and resources (people and time) will be available for marking. In a less than ideal situation for the test setter (large numbers, a heterogeneous group, few resources), texts will have to be understandable and appeal to everyone and items will need to be marked quickly and by few people.

▶ Chapter 13

The question of authenticity is perhaps the most important issue to be considered when choosing texts for tests of reading. This does not mean that only original unadapted texts can be used. A text is authentic if a native or competent speaker of the language feels this is the case. In reality, texts are always being adapted to suit particular purposes without losing their authenticity. *Reader's Digest* can be taken as an example of this. It is rare that

a text which is chosen for a test of reading needs no adaptation at all. However this does not make it necessarily any less authentic.

How much it is adapted or changed will depend on various things, but may involve
- shortening the text,
- removing unnecessary detail,
- making the language simpler, i. e. changing vocabulary or structures,
- making the information simpler,
- correcting any mistakes (typos, incorrect use of language, poor editing of the original).

But it can also involve
- changing names and details so that people, companies and so on cannot be identified,
- changing or removing dates so that the text does not go out of date,
- changing or adding information so that general knowledge or specific knowledge of the topic is not necessary to answer the questions

and may also mean
- adding or removing information to make the setting of questions easier (title, photos or tables),
- adding an appropriate layout to make the text easier to read (and including a title, photos or tables).

When adapting texts, copyright issues must also be thought about. This will depend on your particular situation. It is sometimes possible to write a text yourself to suit your purpose. Depending on how well you can write, these texts need be no less authentic, but it is essential to get a second opinion on this, especially if you are not used to writing the particular type of text.

## 3 What do we do when we read?

Without wanting to go too deeply into what are known as the cognitive and meta-cognitive skills used in the process of reading, it is nonetheless useful to think about what we do when we read, and in particular, what efficient readers do when they read.

**Did you know?**

The efficient reader decides several things within a space of seconds. It is the simulation of these decisions which makes the setting of reading tests complex.

Reading has often been divided into three types, skimming, scanning and reading for detail. Indeed many publicly available exams base their tests of reading on their definition of these.

Recently this separation of skills has been called into question, and in fact, it may be easier to regard the skill of reading as one skill, different aspects of which are applied at different times. It is interesting to apply findings from psychology to test-writing, but here we are only going to consider the most probable types of reading BE learners do and what these mean for test setters.

### Types of reading

In general, it is easier to read if only one reading skill is employed at a time. However this means that learners have to know how to do this. In a foreign language, there is a tendency to only read for detail. There are several reasons for this, including anxiety about missing something important and inability to identify key words, as well as a concentration on this type of reading in schools, where texts are often translated and so detail is important. If you want to test skimming and scanning skills, the learners must not only know that this is the purpose of the test but also how to do it as well as possible.

**Skimming** is also referred to as *reading for gist*. This is done to find out what the text is about, often before deciding to read it in detail, or whether the information needed may be in it (scanning). To be effective, skimming has to be done quickly, so a test of skimming should take this into account. Here are some ways of testing skimming as an isolated skill:

One example of a test format for skimming is matching a short newspaper article to headlines: A number of headlines (2 or 3) can be provided for each text or (more difficult) a number of texts for each headline. Similarly topics and short texts can be used. If you want to test this in isolation, the time for the test must be kept fairly short. The alternatives must be close enough to seem likely but only one should be correct.

It is generally not enough to rely on your own judgement for this but essential to ask others and, if possible, to pre-test the test.

**Scanning** is reading to find particular information. Readers often scan a newspaper before deciding to read an article in detail or may scan a text to find one particular piece of information once the main topic of the article is clear. Scanning is used primarily, and in isolation, for texts which would only be read in detail if they were relevant. These include texts such as job ads, other small ads, descriptions and specifications. They are often not in conventional text format, but consist of lists, tables or diagrams. All the information given is probably not or not immediately relevant and the main purpose of reading is to find a particular piece of information (price, job, times etc.).

One way of testing scanning is to ask the test-taker to identify one text from a range of texts, i.e. find the text in which a certain required piece of information occurs. For instance, the test-taker can be given a number of situations relating to job advertisements and asked to identify which advertisement would be useful for the situation. It should not be possible to identify the text simply by identifying one word in the text or by using features of the text such as layout or pictures.

**Reading for detail** is what takes place once the decision has been made to read the text carefully. As mentioned above, efficient readers can often skim, scan and read in detail at once if necessary. Reading for detail is answering mental questions before reading the text or which someone asks afterwards.

Answering open, multiple-choice or true/false/not mentioned questions on details of the text is the type of reading comprehension test which trainers and test-takers are usually most familiar with, but which is not necessarily the easiest to set. The first step is to identify the details which are important and then to identify the best way of testing these. The type of question chosen will depend on the text.


*In real life a reader decides whether to read a text or not.*

Test-takers will have to know what type of text they are going to be tested on. This will involve the instructions for the test (rubric) and possibly the layout. To make sure the test is as fair as possible, both of these should be included, i. e. tell the test-takers what sort of text they are reading and also let them see this in the layout of the text. Then they will have to know why they are reading it. In real life a reader decides whether to read a text or not. In a test, the test-taker has no choice. However much the instructions tell the reader/test-taker about the purpose of reading, the main reason in reading the text will be to answer the questions correctly and pass the test. It does however make the test more authentic and interactive if a setting and a reason for reading are given. Getting into the habit of doing this also reminds the test setter of these two basic principles of test setting. If you cannot provide a setting and a reason for reading, there is something wrong.

## 4 Instructions for tests of reading

**Remember**

Reading the instructions for a reading test can help you to decide if it would be the right one for your learners. Are the text and the purpose of reading suitable for the target group?

Here are some of the questions readers ask themselves before and during reading:

- What sort of text is this?
- Why am I reading it?
- What information do I need from it?
- How do I get this information in the quickest and best possible way?
- How do I know if I am right?

The most authentic type of test is one where the test-taker knows the reason for reading the text. The most interactive test is one where this purpose, and the skills involved, are as close to reality as possible.

**Example**

Here are some possible instructions for a test of reading:

- *You see the following article(s) in a newspaper. You want to decide which one to read. What is it/are they about? Match the headlines to the articles.*
- *You want to know … Can you find the information in this article?*
- *You want to know … Which article can you find this information in?*
- *You want to know the following information and read this article to find out.*
- *Your boss asks you for the following information. Answer his questions after reading the article.*
- *A colleague has asked you some questions about … First read the questions, then read the newspaper article to find the answers.*

▶ Chapter 14

In general, tests of skimming and scanning tend to be easier than tests of reading for detail. It is not always possible to separate and isolate these reading skills. So it often makes most sense to combine them in one test. This will depend largely on the text type and the purpose of reading.

**Example**

a) *You have received and printed out the following three emails (a complaint, an invitation and a request for information). Which one has information about product CB?*
b) *Now read email 3. What is the customer complaining about?*

So how can we use the answers to create tests? What sort of items can we use to test these skills?

The answers to these questions may depend on other decisions, but in an ideal situation, you can create a test for each of these. The tests could look something like this:

**Examples**

1. Email text taken for learners' area of work:
*Read the emails and decide which of the alternatives to each is the best action to take.*
a) An English answer
b) Forward with instructions in English (or German)
c) Post-it with message: Ring XY about VW

2. A product description taken from the learners' area of work and a description of the product features the learner needs:
*Is this product suitable for your needs or not?*
*If not, why not?*
a) It is too big.
b) It is too expensive.
c) It is too small.

3. A marketing text with multiple choice questions on the most important points.

All these assume that

- test-takers have the same level of knowledge of the subject, i. e. are a homogenous group. Clearly this knowledge is going to play a big part. Emails taken from one sector will not be intelligible to another nor will it be always possible to answer the question on the course of action. A test of this nature becomes meaningless if the course of action necessary is a matter of opinion. The same would apply to the second situation.
- being able to answer the questions on the third text will depend on firstly some knowledge of marketing, secondly the time available, and thirdly on understanding the questions.

## What are the best sorts of questions to ask?

If you take all the information you have on the learners' reading habits into account, then there may be a large number of types of questions which can be asked, but if you want to set or choose a test which can be used again or (with the same format) can be used for a number of different test-takers and tests reading as far as possible in isolation, you will probably have to think about multiple-choice, true/false/not mentioned or short answer questions (SAQs).
Whichever of these you use, watch out for the following:

- Double negatives in questions make them difficult to answer.
- Do not test mathematics.

**Example**

*The survey found that 65 % of customers prefer blue as the colour for household detergents, 25 % pink and 10 % orange.*
The question *What proportion of customers do not prefer blue detergents?* not only involves a negative but also involves adding 25 and 10, admittedly not difficult but unconnected with understanding the sentence.

- Do not make the questions more difficult than the text.
- Do not give unintended clues or trick the test-taker into giving the wrong answer.
- You do not want to test whether a learner can copy the right answer from the text, but you must also be careful not to make the questions more complicated than the test.

**Did you know?**

In a test, most people read the information and do the questions in the order in which it is presented on the page, even if you tell them they can do otherwise. Only experienced or well-trained test-takers do different parts of the test in a different order from that in which they are presented.

- Does the text give you enough questions to give you a meaningful result?
- Is the task related to what we do in real life?
- Are all the questions independent of each other, i.e. does getting the answer to one question wrong mean that the test-taker will make more mistakes later on?
- Are the instructions clear?
- Are the questions in an appropriate order? This will depend on what you are testing. In a test of reading for detail it makes sense to have the order of the questions in the same order that the information appears in the text, but this would probably not be suitable for a test of skimming or scanning.
- Think about the layout of the text, the instructions and the time the test should take.
- Where do you place the questions? It is not always logical to read the text first and then answer the questions. In general it is most appropriate to place general questions before the text and questions of detail afterwards, but it may be best to always have the questions first and in the instructions ask test-takers to read the questions first. This also gives them the reason for reading.

## Other question types

Publicly available exams offer additional formats, some of which take a considerable time to set and many of which also test grammar, vocabulary and understanding of text structure and cohesion. To make these valid and reliable, extensive pre-testing is absolutely essential. Two other examples of reading comprehension test formats are:

### Putting paragraphs in the right order

This is difficult to construct because the test setter must be certain there is only one possibility. This is not always the case.

### Sentence completion

The candidate has to complete part of a sentence using information from a text. Here more than one answer may be correct but the number of correct answers should not be too big. A record must be kept of what is considered acceptable.

You may also wish to think about whether test-takers should be able to use a dictionary or be allowed to ask the meaning of words during the test. This will depend on how closely you define what you are testing. There are reasons for and against giving them some form of help but this means that the results will be less reliable as other skills are involved.

▶ Chapter 9

■ In general tests of reading should differ from tests of vocabulary in that getting an answer right should not depend on knowing one particular word. This would be better tested in a vocabulary test. However it is always the vocabulary load which makes tests of reading comprehension easy or difficult. Teaching learners to deal with different texts in different ways is a classroom task. The test of reading should aim to see whether they have learnt the relevant skills and can use them to understand the meaning of suitable texts. ■

**To think about:**

- What sort of texts do you need?
- What sub-skills (vocabulary, grammar) are involved in understanding the text?
- What sorts of questions should you ask?

CD-ROM

*"In the case of world Englishes and cross-cultural communication it seems clear that the comprehensibility of speech cannot be reduced to a particular accent or variety."*

Van der Walt, Christa (1999): Business English and International Comprehensibility. In: Business English: Research into Practice. Longman, p. 49

# 8 Listening

## Listening, listening again and finally understanding – but how do we know?

- What is listening comprehension?
- How do you choose listening texts for tests?
- What sort of questions should you ask?

Testing listening raises all sorts of questions in addition to those already mentioned in the chapter on testing reading. There are several similarities between testing reading and testing listening, but as well as considering what type of test you are setting and thinking about what types of items you are going to use, you will need to think about several more things as far as listening is concerned. Although each skill is complex in itself, several questions, including that of authenticity, are more important in the case of tests of listening comprehension. In many ways, listening can be compared to reading, but not all the ways in which reading comprehension can be tested are suitable for listening. There are several reasons for this, the main one being that test-takers have no way of returning to the text. Others will be considered in this chapter.

### 1 What is listening comprehension?

There have been different ideas on what listening comprehension means throughout the history of EFL teaching. Many of the approaches arising from these are still popular today, perhaps especially as testing techniques. This does not mean that they are particularly valid in terms of BE. They are probably still used because they are practicable, i. e. easy to use in tests.

**Did you know?**

Tests of how well test-takers know and can use the phonetic alphabet were also considered tests of listening for a time, but while the phonetic alphabet is still used in schools, a test involving the phonetic alphabet can be considered a test of knowledge rather than one of language skills.

The most obvious, and perhaps extreme, example is dictation, still sometimes considered a test of listening. In a dictation, test-takers hear a spoken text (usually read aloud) and write it down. More often than not, they are awarded points for correctness of spelling and punctuation, usually absolute correctness.

Writing a dictation may be a test of listening, but it cannot really be considered particularly valid or authentic as it says nothing about how well the test-taker can understand the communicative force of the message. If this is not important, then there is no reason why a dictation should not be used as a test. If you want to test orthographical accuracy (spelling and punctuation) then a dictation may well be a valid and practicable way of doing this.

Whatever form of listening comprehension text and test are used, the following will apply:

- Most test-takers feel that tests of listening are more difficult than tests of reading, although the results do not reflect this.
- Memory has to be taken into account. In general a language test aims to test language skills rather than any other skills such as memory. This means that in a test of listening, only short-term memory can justifiably be tested, as long-term memory depends on a range of factors unconnected with language ability.
- You will have to decide how many times test-takers can hear the text and whether there will be pauses between parts of the text for them to answer some of the questions.
- The questions must be in the same order as the information occurs in the text.
- You will have to think about whether you allow test-takers to take notes while listening. In listening tests it is often more appropriate to have tasks which involve some activity on the part of the test-taker, i.e. filling in a diagram or following a map.
- Test-takers can be doubly penalized if they miss one question, i.e. by not listening for a moment, or if there is a disturbance or additional noise in the room where the test is being taken. If they miss a question, they may get the questions which follow it wrong, because they do not know where they are in the recording.

It is probably mainly because of these issues that BE trainers do not often set tests of listening, but instead concentrate on other skills when designing tests. However, whether or not you set your own tests, there is no doubt that listening is an important skill and one which is a definite part of BE courses, course books and business life in general. This means that it is worth thinking about how it can best be tested.

The reasons for testing listening skills will be that you want to find out how well your learners actually can understand what they hear, how much they understand and to what extent, i.e. how high their level of listening comprehension is.

## 2 People or equipment?

Before we come to types of tests of listening comprehension, we should however think about how the text is transmitted, i.e. by humans (speaking or reading aloud) or by using equipment.

Some of the very early UCLES (at that time not yet Cambridge ESOL) exams had a dictation test in them, read aloud by a native speaker. When the notion of what listening ability and comprehension meant began to change in the 1970s, the test changed to one of test-takers answering questions on a text or texts they heard. As technology was not readily available in all the countries in which the exams were taken, for many years, from

1970 until the early 1980's, this test of listening comprehension was conducted by native-speaker examiners who read the texts aloud twice, with pauses the second time. This was clearly a compromise for reasons of practicability, but also must have affected the choice of text, as presumably the texts used were those involving only one speaker. This must naturally have limited the range and type of text chosen.

This method of testing listening comprehension, while clearly a practicable one, is nowadays usually rejected because it is felt to be inauthentic. The view of listening comprehension has changed from a classroom-focused definition (just listening to a text and writing it down correctly) to a communicative one, i. e. listening comprehension as a form of (mental) interaction with what is heard in order to make some sort of use of this. This affects the following issues:

- Authenticity of tests of listening comprehension
- Listening skills tested
- Instructions to test-takers
- Choice of text
- Type of questions

All these things mean that there are very few tests which in real business life are read aloud (and this more than once) by one person to a group of others. This method of testing listening comprehension can only be regarded as an emergency measure when there is no equipment available for producing or playing a recorded text.

If you really feel you have to read a text aloud in a test of listening comprehension, make sure that you have thought about it very carefully, perhaps by asking yourself the following questions:

- Would this text be read aloud in real life?
- Why?
- Who would read it?
- What sort of accent would the speaker have?
- How fast would he/she speak?
- Would there be any background noise?

## 3 Authenticity

If we are listening to a human being speaking rather than a recording, we may have the chance to interrupt the speaker to ask for clarification or repetition. If this is not possible, i. e. if it is a lecture, we can take notes or ask questions afterwards. Making tests of listening comprehension more authentic inevitably blurs the distinction between the different activities and to what extent these are being tested alongside listening comprehension. Are we just testing listening or, as in the case above, also note-taking, asking questions, interacting? How do we assess and score these appropriately?

Using recordings usually precludes the possibility of interrupting to ask for clarification, however authentic this may be. We cannot interrupt a recording without missing something. In a real life situation, we may ask someone else who is also listening to explain what is being said to us, but at the risk that we will both miss the important information. We may not be able to hear it a second time. There are of course cases where the message is given again (announcements are often repeated later, the news is repeated at regular intervals) or we may be able to confirm what we think we have understood in some other way (by referring to printed information or asking at an

**What is it?**

The word *recording* will be used for all listening texts transmitted through equipment, whether they are intended to simulate text which is recorded in real life (e. g. radio broadcasts) or spoken text which is not recorded (e. g. conversations overheard at a conference) or live texts (e. g. announcements at a conference through a PA system) or not.

information desk, for instance). In order that a test of listening comprehension is as authentic as possible, all these should be taken into consideration.

All this means that a test of listening comprehension, if it is to test exclusively listening comprehension, is bound to sacrifice some authenticity for the sake of validity and practicability. Most tests of speaking involve a certain amount of listening comprehension. However that is not their main focus and they are approached, set and marked in a different way to tests where the main focus is on whether the test-takers have understood what they have heard. It is this second type of test which is the focus here.

*Is this authentic?*
*Overhearing other people's conversation*

## 4 Listening skills

Bearing in mind the features specific to listening, the three main listening skills are basically the same as those in reading comprehension:

- Listening for gist
- Listening for detail
- Selective listening

These can be explained by using a typical news broadcast followed by the weather and traffic reports. Listening to this probably means that the listener applies all three skills. In the news itself, detail is generally important, i. e. who died, how, where the earthquake was and what the results of the election were. If we think of the questions we ask after the news about the parts we may have missed, they will be questions of detail rather than otherwise. The weather report for one particular area is a slightly different matter, as here it is the general picture which is usually important, and we get this by understanding the general meaning of all the various parts. The national weather report is usually an example of selective listening followed by listening for detail, i. e. once the area has been identified, the details are important. Listening for gist also involves identifying the opinion or attitude of the speaker(s). Whereas this may be done by identifying particular words used, it is more often a case of identifying and combining several things the speaker says to get the overall picture and thereby understand the attitude, opinion or stance of the speaker. In terms of the weather forecast, this may mean understanding words like *bad* rather than *drizzle*.

Listening to and understanding the traffic news effectively is a matter of selective listening, as we generally completely filter out or ignore the reports from areas or roads we have no intention of driving on. We only listen for those roads which are a part of our route. Having made this initial selection, however, we then generally listen for detail. Like reading, listening ability consists of applying the right skill at the right place and time. The difference between reading and listening in this respect is that if we apply the wrong skill to listening, we have little chance of correcting ourselves and therefore are not able to achieve our aim, that of understanding the information we need.

**To think about**

You may argue that the listener usually decides what he or she wants or needs to understand and remember. However this can be legitimately decided by the test setter as long as the test-takers know the reasons why they are listening.

## 5 Instructions

Clear and detailed instructions which are given to test-takers at the right time are extremely important in tests of listening comprehension. Although we do sometimes hear something and then spontaneously decide to start listening to it, there is no good way of testing this fairly, i. e. without putting test-takers under a great deal of stress. In practical terms, this means that test-takers must know what they are going to hear, why they are listening to it, how many times they will hear it, what questions they will have to answer, how much time they will have for this and in what form they should give their answers. They should know all these things before they start listening. All these things are important in all tests, but the transient nature of listening makes them more important in a test of listening comprehension, if it is to be valid, reliable and fair.

▶ Chapter 14

Instructions for tests of listening comprehension are often recorded and printed so that test-takers hear them and follow them on the page at the same time. This is not absolutely necessary, but may help test-takers to get used to listening to recorded voices. The instructions should also mention the pauses between tests and texts and the time allowed to read the questions. In the professional recordings made for international exams, these are included in the recording so that all candidates have exactly the same amount of time to read and answer the questions. If you are setting the test yourself, you will have to decide what an appropriate length of time for this is. This will depend on the level of the test, the amount of reading which needs to be done beforehand and the number of questions to be answered. Too little time is clearly unsatisfactory, but too much time can also be frustrating and counter-effective. Remember test-takers cannot use any additional time to check their answers, as they can in other tests.

There may be a time when each test-taker will be given their own listening device with the recording on it. That would change the nature of testing listening entirely and raise the question of whether it would be authentic to be able to rewind/listen again as often as you like. This would depend on how listening skills are defined, for instance how much a speaker in a face to face conversation should be prepared to repeat and clarify and how many requests for this will affect the view of the listener's language ability and the overall communication.

A note on recordings and questions in BE course books: although the recordings are often highly authentic, professionally recorded and interesting, the tasks accompanying them are usually intended for classroom use and so are not always suitable for tests. They are often designed for classroom activities such as discussion. This means that if you use a recording which is already available and which you believe to be suitable, you will probably have to design questions on it to use in a test. This will also apply to recordings from other sources, such as radio broadcasts or podcasts.

**Remember**

Depending on what your test is to be used for, you may have to clear up copyright questions relating to the use of broadcasts.

## 6 Texts for tests of listening

### Choosing a text

Choosing a text can be approached from two angles. You may wish to choose a text which is authentic for your learners and then decide which listening skill you wish to test. This will mean finding out what sort of texts your learners need to hear/listen to and understand.

The word *listen* is more often used in tests than *hear* as we can safely assume that test-takers are listening consciously rather than hearing something by chance, as they may do in real life. This distinction can be used in the instructions for the test or task:

**Example**

*You are attending a conference on marketing techniques. During the coffee breaks you hear several announcements. Read the questions below. There is one question for each announcement. Then listen to each announcement and answer the question. You will hear each announcement only once.*

Choosing a text also means thinking about why your learners will need to listen to the text, i.e. what you are going to test. In many cases a recording is "just" interesting, i.e. leads to discussion and provides new ideas for this, but this may mean that it is not suitable for a test of listening comprehension.

The other way of choosing a text or texts for a test of listening comprehension is to decide which skills are important for your learners, which skills you have been concentrating on in class or which skills they need to be tested on, for instance for placement or diagnostic purposes. Having decided on whether listening for gist (the general idea) or detail is important or whether it is a question of selecting the important information, you can then go on to find a text which will be suitable for this. Here are some ideas of which texts are suitable for which skills:

**1 Listening for gist**
- Largely non-factual radio programmes on business topics
- Business discussions (for questions on opinion and attitude)
- Non-factual speeches, lectures and talks where a general picture or an opinion is given

**2 Listening for detail**
- Business news broadcasts
- Information programmes
- Lectures on facts and specific information
- Announcements (e.g. from railway stations, conferences, radio reports)
- Business meetings on specific issues where details are discussed
- Presentations on specific topics

**3 Selective listening**
- Announcements from railway stations, conferences, radio reports (with different tasks to those used in tests of listening for detail)
- Announcements (travel, conference)
- Traffic reports
- Travel information

**The following will also have to be considered in a test:**

- Length of text
- Number of speakers
- Accents and source of texts
- How many times
- Background noise
- Test environment

It may seem that we do less "pure" listening (i. e. only listening) nowadays but often also watch something while we are listening. This of course applies to real life situations as well as the classroom. However there has been very little use of video material in tests so far for any real judgements to be made about it.

## Length

Texts should not be made artificially longer just so that more questions can be asked on them. For tests of listening for detail, texts will tend to be longer, and for selective listening, usually shorter. Tests of selective listening usually consist of several texts, with one (or maybe two) questions for each text. All the texts may be in one context or several contexts may be used. Remember to take the test-takers' memory and concentration into account. If a text is very long but otherwise suitable, it may make sense to break it up into logical parts so that test-takers' memory is not overloaded. Texts in which there is a change of scene or time or event are more difficult to follow. Changes should be clearly indicated, perhaps by a pause in the recording or instructions outside the recording itself (e. g. "Now listen to the next part").

## Number of speakers

If there is no visual support for the listening text and it is important to identify the speakers in order to answer the questions, then the test-taker must be able to do this easily. For this reason, there are often only two speakers in one continuous text, a man and a woman. More than three speakers in a recorded discussion makes it very difficult to follow, and therefore, although more authentic, less reliable. Those test-takers who can identify different voices easily will have an easier task than those who rely on visual input to identify speakers. Not only the number of speakers but also their relative roles should be considered. Three speakers who are all colleagues will be more difficult to differentiate between than a boss and two colleagues who are members of his staff.

## Accents and source of texts

As much of the communication in BE is nowadays between non-native speakers, it may be more authentic to use examples of non-native accents. Whether you use recordings with non-native speaker accents will depend very much on your view of language and your learners. It is fairly common practice in international exams. If you decide to use non-native accents, this will also clearly affect your sources. In tests of listening it is best to establish the context before the test-takers hear the recordings, and it is also often a good idea to indicate who the speakers are and where they come from. This will all serve to make the text more authentic. Also remember that natural spoken language consists of redundancies and pauses. Leaving these out not only makes texts less authentic, it also makes them more difficult to understand.

### Did you know?

Some say that 80 % of all business interaction takes place without any native speakers present, others feel this figure is exaggerated. It is mentioned by Jürgen Beneke in "Englisch als Lingua franca oder als Medium interkultureller Kommunikation?"

*In: Grebing, R. (ed.) (1993): Grenzenloses Sprachenlernen. Festschrift für Reinhold Freudenstein. Cornelsen, p. 54–66*

### Background noise

Background noise and interference are of course authentic, but depending on the type of text and level of test-takers, may impose an unnecessary burden on them. However, any test involving telephone conversations should simulate this as authentically as possible.

## 7 Other factors

### How many times?

This will also depend on what skill you are testing and how authentic you wish your test to be. Once is usually enough for tests of listening for gist, whereas in situations where a second chance would be possible, recordings can be played twice. This also relieves the load on test-takers' memories. If the text is very long, then it can be played once in entirety and the second time in parts. If one context (e.g. a conference) is being used for several texts and these are to be played twice, then the second time should immediately follow the first. This is even more important if the context changes from text to text. Playing a recording more than twice is unusual and should not be necessary. A text should be played twice if the questions are very close together. In general it is not a good idea to have different questions on parts of the text which are very close together, because if test-takers miss one answer they may miss the next one too. For this reason, it may be better to use several short texts and play them only once rather than use one long text.

### Test environment

This means the room where the test is to be taken. For a test to be fair, all candidates must take it under the same conditions. For a test of listening comprehension, this means using suitable equipment to play the recording in an environment with little background noise. Before you start the test make sure that everyone can hear the recording well, adjust the volume or rearrange the seating.

## 8 Test Questions

### Suitable question types

**For tests of listening for gist:**

- Open questions
- True/false
- Choose a text/title

**For tests of listening for detail:**

- True/false
- Note-taking
- Open questions – single word
- Gap-filling

**For tests of selective listening:**

- Multiple-choice questions can be used for these tests if the stem and the distractors are not too long.

Multiple-choice questions are not suitable for all tests of listening if the options contain a lot of text and information, as test-takers have to remember a lot to eliminate distractors and decide on the right answer.

## What should the questions test?

Test-takers can be asked to understand gist, main points and details, both of what one speaker says and of the interaction, i. e. the combined message between two speakers. A test of listening should not simply be a recorded version of a test of reading but should aim to test some of the specific features of listening to spoken language or interaction. As the meaning of spoken language depends on intonation, test-takers can also be asked questions on attitude or intention, as long as the answers to these are not too subjective.

■ Bearing all this in mind when doing listening activities in BE classes will help you to develop learners, listening skills as well as to assess course book listening material as to whether it is suitable for your learners' purposes.

**Remember**

There is no perfect test!

Varying the type of listening your learners have to do in class and differentiating between the different skills they apply will help them in everyday business life. If you do decide to set or use a test of listening comprehension, then taking all these into account as far as possible will not only mean that your test fulfils the criteria for a good BE test but will also make it as fair as possible for test-takers. ■

### To think about:

- Where can you find suitable texts?
- How can you make a test of listening as authentic as possible?

CD-ROM

*"Students write to accomplish a variety of purposes and use a number of different genres to do so. Purpose in writing determines the nature of their writing."*

O'Malley, J. M. / Valdez Pierce, L. (1996): Authentic Assessment for Language Learners. Longman, p. 137

# 9 Writing

## Authenticity, message and accuracy

- What is an authentic test of writing?
- What do you expect?
- What are the main features of tests of writing?
- Who will mark the tests?
- How accurate should writing be?

Most writing tasks for BE are fairly easy to set and most BE trainers will have little difficulty thinking of suitable tasks for writing which can be used in tests. Writing is probably the skill which is easiest to test authentically and directly. Although some of the enabling skills of writing, like vocabulary knowledge and grammatical accuracy, can be tested in closed-item formats such as multiple-choice, writing itself can best be tested directly, i. e. by getting test-takers to write something. If what they are asked to write in a test corresponds to what they have to write in real life, the test task will fulfil the criteria of validity and authenticity. If all test-takers are asked to carry out the same or a similar writing task, the test may be reliable as far as the task is concerned.

However what it will not be (yet) is objective or practicable nor will the results be reliable. Without marking criteria and some idea of what is expected, what test-takers write cannot be marked objectively and will, in most cases, be marked quite differently by different raters. This is why setting a test of writing involves more than just the task; it involves thinking about what is expected and how it should be marked, scored and graded. The issues of marking, scoring and grading are of particular relevance to tests of writing.

▶ Chapter 13

## 1 What is an authentic test of writing?

The first thing to think about is the task to be set. Most writing in BE is probably of emails as well as reports and maybe business letters. However, whatever types of writing are being tested, the same principles apply to most tests of writing. Although, as already mentioned, tests of writing are fairly easy to make reasonably authentic, there are some differences between writing in real life and writing in a test and these should be remembered. Some of these are:

- The time constraints involved, as a test usually has to be completed in a specific time.
- Planning and revision are two processes which form part of authentic writing and yet are rarely included in tests. This is usually for reasons of practicability. Much planning and most revision of what has been written does not take place immediately before or after the actual writing but at a later stage, after a few hours or even a day or two. This is clearly almost impossible to achieve in a test and so can only be simulated. However it should be thought about when setting the task and especially when deciding how accurate and linguistically correct the language produced by the test-takers should be.
- Tests of writing usually involve test-takers writing by hand, something which is rarely done nowadays. This is also a question of practicability. Cheating is one of the issues test administrators have to deal with, and one form of cheating is clearly unauthorised access to and use of resources, such as the Internet. Allowing test-takers to write electronically makes this access easier and involves additional supervision or electronic restrictions to outside access.
- The people who mark writing tests (raters) are usually teachers or trainers and have different expectations of the piece of writing to those who would normally read it and for whom it is intended.
- There is usually some access to reference when writing under normal circumstances, even if this is only in the form of an electronic dictionary or spell-checker. The more competently a person can complete a writing task, the less reference or help will be necessary as more sub-skills and enabling skills will be automatic.

▶ Chapter 14

Some tests allow the use of dictionaries, but there are several things to be considered before doing this. There are many advantages and drawbacks to letting test-takers use dictionaries in tests. It is clearly more authentic to allow them to look up a word they do not know, but it means that they should know how to use a dictionary properly and possibly under time pressure. There are often practical problems involved, such as who provides the dictionaries, should they all be the same type (monolingual or bilingual). In general, in short tests of writing, where time is a factor in the test, there are more arguments against than for allowing the use of dictionaries. Long exams, such as 4-hour school leaving exams, where content is also marked, often allow the use of dictionaries, but the testing situation in those cases is quite a different one from that in BE. There is also a danger of cheating (if looking things up is seen as cheating). Test-takers could stick pre-prepared lists of phrases, vocabulary etc. in the dictionaries. Checking each dictionary involves a great deal of administration, as does only allowing dictionaries provided by the testing institution.

▶ Chapter 14

Some constraints applying to tests of writing (as opposed to tests of other skills) are:

- Tasks have to be set which give test-takers the opportunity to produce pieces of writing which show what they can write and can also be easily assessed. It may be worth considering setting two different tasks in a test of writing so that not too much emphasis is placed on one topic and style of writing. Having more than two tasks in one test of writing usually takes up too much time and leads to repetition.

▶ Chapter 13

**Remember**

Handwriting, for instance, may be considered a non-linguistic aspect of writing, and not something which a BE trainer should teach, test or assess. However if handwriting is impossible to read, there will be no way of judging any other parts of the writing.

- There are no correct answers in tests of writing. Decisions on what is acceptable and can be awarded high points and what is unacceptable will inevitably be partly subjective but will always have to be made, whether by one person (the worst case) or by a group. It will also have to be ensured that raters understand and stick to the decisions made on criteria.
- Some aspects of writing are linguistic and others are non-linguistic. The following are all aspects of marking but cannot all be easily classified into linguistic or non-linguistic: handwriting, style, layout, structure, spelling, ideas and arguments, ease of reading, length of sentences, paragraphing, rhetorical features, and more. BE trainers may only wish to test and rate the linguistic aspects of writing. However separating these from other aspects is not always easy and will have to be specified and described in the marking criteria. This may mean stating clearly what is and is not being marked.
- Raters have to be competent in the language concerned. Extended writing cannot usually be marked automatically or by using clerical marking. Language competence is, however, only one of the qualifications for raters.

There are tests of writing which are carried out electronically. It is also possible to mark these automatically or semi-automatically. This is something which at present can probably only be done by large testing bodies with the necessary financial and technical resources. One example of such a body is Pearson Educational which has introduced artificial intelligence-based, automated marking of English papers for the exam PTE Academic. This has led to some concern about the nature of the writing which will be produced and whether the marking will be as reliable as human marking, but at present it is not really possible for most non-experts to judge. Pearson claims that its system marks as accurately as human markers can (and of course will not suffer from moods or tiredness.).

Apart from the constraints involved in tests of writing, there can be little or no excuse for setting tasks which are not authentic. This applies in particular to BE, where the range of writing tasks is fairly limited and particular tasks easy to identify. Just as language users never read without a purpose, they also never write without a purpose and the main intention of writing will be to communicate something effectively. This involves more than just correct language, so it would be a mistake to mark writing purely on the basis of grammar and vocabulary. These can be tested just as well, if not better, using other types of tests, such as multiple-choice or gap-filling items.

**To think about**

Letters (snail-mail) are becoming less common but are still being written. However, there is no point in testing the writing of letters if the test-takers do not need to write them. Likewise there is no point in testing the conventions of letter writing if these are prescribed by the company, and all the writer has to do is follow them. The same things apply to faxes.

Almost anything which a learner has to write for work purposes can be transformed into a test task fairly easily. The process involves answering much the same questions as when choosing teaching material:

- Why is the writing taking place?
- What form is it in?
- What are the main features?
- What register is used (formal, semi-formal, informal)?
- Are there conventions for this type of writing?
- Are these general, national or company conventions?
- How is the message best conveyed (directly, indirectly)?

These questions are probably answered in the teaching material you are using and will need to be considered when developing the marking criteria for a test of writing.

The following questions, which apply particularly to tests of writing, require more time and thought to answer:

- What do you expect (in as much detail as possible or with one or more examples)?
- What input will be given?

- What form will the input be in?
- How much detail will there be in the input?
- What instructions (besides the input) will be given to the test-taker?
- How much should the test-taker write?
- How much time will the test-taker have?
- What are the marking criteria?
- Who will mark the tests?
- How accurate should the piece of writing be (how many grammar, vocabulary, spelling and punctuation mistakes are acceptable and how do we score them)?

## 2 What do you expect?

Deciding on your expectations means establishing a level both for the task itself and for what is written.

Defining the level of difficulty of a task is not easy, as most tasks can be completed in different ways by test-takers at different levels. Nonetheless there are tasks which can only be completed at or above a certain level of language competence. A task for all levels could be to write an email inquiring about special services in a hotel.

Depending on the level of the test-takers, the email could get full points for
a) conveying (parts of) the inquiry but with mistakes of style or language,
b) conveying the complete inquiry with no or few mistakes in style or language,
c) conveying details of the inquiry with several mistakes or
d) conveying details of the inquiry in an appropriate style and with very few mistakes.

Even an attempt at a simple categorisation like this raises several questions, such as: When is the inquiry considered complete? What is a mistake in style? How many mistakes are several?

### Tasks

Depending on the level of the learners, a BE task could be within a range like the following:

- A standard email replying to an order, offer or invitation where no difficult issues (price increases, negotiating an offer or refusing an invitation) are involved
- A more difficult email such as one making a complaint, raising an objection to something, refusing an invitation
- A formal email or letter which may have particular (e.g. legal) implications for the individual or the company

Having decided what tasks test-takers can be expected to complete, expectations of what they write will have to be established. Which pieces of writing get full points and which do not? A very common mistake is to award only "perfect" pieces of writing full points. "Perfect" usually means with no grammatical, vocabulary or spelling mistakes (sometimes it means only this). However, almost any piece of writing, even if done by a competent native speaker, can be improved in some way. The constraints on testing writing make it unrealistic to expect perfection of any sort in something written in a test by a learner of English. What is considered right or wrong will always be within a range and will always involve a certain element of subjectivity.

**Remember**

Just as there is no perfect test, there is also no completely objective test, as all the decisions about the test from the construct to the items and all the other details will be decided by one or more people. The more different views are taken into account, the less subjective and therefore the fairer the test will be.

▶ Chapter 10

**Did you know?**

Examples of pieces of writing at CEFR levels are provided by the Council of Europe.

Decide on the best you can expect from the group of test-takers. One way of doing this is to get a class to do the test and select the best piece of writing from those produced. Decide if this piece could realistically be improved and yet still be within the range which a member of the target group of learners could produce. If not, this piece can provide you with the benchmark for a full score.

Full marks should be given to the best writing which can be produced by the particular test-taker under the circumstances. Similarly you will have to decide what constitutes a " pass, i. e. the minimum. Once you have decided on these, you can go on to discuss the reasons for the decision and what aspects of the piece of writing you want to focus on when marking. There is no single right answer here, but it is worth considering what a test of writing tests specifically and what can better be tested in other tests. The smaller the number of criteria you have to consider the better. Spelling and punctuation, for instance, can only really be tested in tests of writing, so if you think these are important, then they should be included in the criteria.

## 3 Input and instructions

Writing tasks can consist only of instructions or can also be based on some form of input.

### Input

The input can consist of the situation, the setting, the type of writing (email, letter, report, etc.), the roles of the writer and the person addressed. There may also be input (something for the test-taker to respond to in writing). This may be written, whether or not in the same form as the result expected, or in the form of a diagram or picture. The input can even be a longer text or a recording, although this makes the test less one of just writing and more a test of integrated skills. This makes it more complex to mark as the influence of the other skill(s) has to be taken into account. Unless only pictures are used and there are no instructions or these are given orally, reading will always play a part, however minor, in any test. This is why language tests for illiterates are very challenging to construct.

**Example**

| Input | Expected form of response |
|---|---|
| Email request | Email response |
| Job advert | Letter of application |
| Graph | Written report |
| Lengthy report | Summary |
| Article | Email reaction |
| Job application | Standard letter of rejection |

If you want to test writing as far as possible in isolation, then the task or the input will have to be kept to a minimum length and use simple language. The task should not be worded in language above the level of the test-taker but neither should it be possible for the test-taker to lift sections from the task or the input. This is something which is also done frequently in real life but makes the results of tests of writing less reliable. If you do not know the test-taker, you cannot possibly know if he/she could have written the piece on his/her own, and if you do know the writer, your judgement will be subjective and possibly unfair. For these reasons it is best for the input or task to be formulated in language which would not help the test-taker to complete the task.

**Remember**

It is not fair to penalise test-takers for bad tests.

One way of doing this is to give instructions to the test-taker, e.g. *write a reply and say when you are arriving, refuse the invitation, give reasons, specify prices* etc. Another way is to use input which is not of the same format or register as what you expect the test-taker to write, e.g. an article on a website you want to send to a colleague and give some information about or a rude complaint from a customer which has to be dealt with politely.

### Instructions

The instructions are distinct from the setting and the input. Their intention is to inform the test-takers of what they have to do in the test, e.g. *Write an email in response to the email given here. Write clearly and do not write more than 45 words. You have 15 minutes to do this.*

You will have to decide what you want to put into the instructions. Although this applies to any instructions for tests, a test of writing involves more parameters than most and so the instructions have to be thought about carefully. On the one hand, you do not want to impose a huge reading load on the test-takers and possibly alarm them with a large number of detailed instructions, but on the other hand, it may be important to remind them of certain things. Taking the above as an example, *write clearly* is considered by some test setters to be a superfluous instruction whereas others feel it is important to remind test-takers of some basic principles. The same applies to instructions such as *Decide first on your arguments* or *Read through what you have written when you have finished writing.*

▶ Chapter 14

## 4 Other factors

### How much should the test-taker write?

Whether you want to restrict the test-taker to a certain number of words will depend on the task. The number must be a realistic one for the task. If you specify a number of words, then you will also have to decide whether it is a minimum or a maximum number and how you will mark pieces of writing which are longer or shorter than this. The decision on the number of words depends on the task you set and can only be made by looking at authentic examples of the same sort of writing and seeing how long these are. Whether you put the number of words into the instructions will depend on how important the length of the piece of writing is for the completion of the task. You may not want the test-takers to write too much (in an email for instance, as this could obscure the main message) or too little (as this may not give you a good enough picture of their writing ability). On the other hand, length may not matter for the completion of the task, so you may not want to specify it at all. All these decisions will have to be made before the instructions are written.

### How much time will the test-taker have?

Closely linked to the length of the piece of writing is the question of time. Unlike tests of other skills where the time is dependent largely on the length of texts and the number of questions and cannot be extended by the test-takers themselves, in a test of writing test-takers (theoretically, at least) continue writing for as long as they want.

Deciding on the length of time you want the test to take is best done by starting with the task. What task do you want to set? How long should it take in real life to do this? A short email will take less time than a lengthy report. Try the tasks out in class and allow about

10 minutes more for reading the instructions and familiarisation with the input at the beginning as well as checking and proof-reading and possibly counting words at the end. Do not allow too much time as this can lead to boredom and frustration (especially if the test-takers have to wait until everyone has finished) but do not expect test-takers to do a real-life task in less time than can realistically be expected of them. You may, however, wish to ensure that test-takers can write to a certain standard under time pressure, a skill sometimes important in business.

**To think about**

Input such as an email with the task: *Reply to the email* is less specific than guiding points for the email such as:
- Thank the supplier for the order
- Explain that there will be a delay
- …

The following will influence the time you need to allow for a test of writing:
- Long, complex input will take time to read and digest.
- Specific details of what to write or guiding points will take time to read but make the task of writing easier as the test-takers will not have to think of these themselves.
- Specifying a minimum number of words means you must be sure that this number can be written in the time. Also think about and, if necessary, say whether you want the test-taker to count the words and give the number, and take into account that this in itself will take some of the test-taker's time.

Also decide how you will mark tests which have not been completed in the time allowed.

## 5 Criteria

▶ Chapter 13

Only the particular issues connected with tests of writing are considered here. Some have been mentioned above – handwriting, style, layout, structure, spelling, ideas and arguments, ease of reading, length of sentences, paragraphing, rhetorical features. The first step will be to decide which of these are important enough to give a mark in a test of writing for BE. This can be discussed with learners and their own ideas of what is important in writing used to develop criteria. Remember that dealing with more than 10 (at the most) criteria means a lot of parallel thinking and decision-making if they are to be dealt with fairly and in a short time. Although the criteria will vary depending on the business context and the type of writing, the main criteria for writing in BE will most probably be general comprehensibility of the message, content and inclusion of details, appropriateness of style, correctness of grammar, vocabulary, spelling and punctuation. You may wish to include all or some of these and possibly others.

Define each of them and make sure that there are no overlaps between the criteria. For example, how much do mistakes in grammar interfere with the overall message? Make sure you do not penalise test-takers twice for the same thing. Having clearly defined and separate criteria will make it easier to mark different aspects separately and not penalise test-takers twice for the same shortcomings.

How to weight the criteria will be the next decision, whether you want to give equal (i.e. the same number of) points for each criterion or whether you wish to place more importance in terms of scores on, for instance, style than on grammar or spelling.

The next decision to be made will be how to define different levels and scores within one criterion. You may have the five criteria above with 5 points for each, leading to a possible total maximum score of 25 points for the whole writing test. Each criterion may score 5, 4, 2, or 1 point. You may also want to include 0 points (even though possibly no-one will get this). You may also want to make one of the criteria a knock-out criterion, for example if the task has been totally misinterpreted, then a piece of writing is often awarded no more than 0 points, even if it is linguistically perfect.

**Example**

Play at being the most awkward test-taker imaginable and try to deliberately misinterpret a task, but in such a way that it could still be considered valid. This will ensure that so-called "misinterpretations" of tasks are not always automatically given 0 points. If the task can be seen in a different way to that intended by the test setter, then either the task should be changed or that interpretation accepted.

**Did you know?**

When setting scores, it is better to have even numbers of stages as generally raters will otherwise tend to choose the middle one.

0 points will also have to be defined. 0 points can be given if nothing is written at all, if what is written is totally incomprehensible or if it has nothing to do with the task set. Although this can happen in high-stakes exams or if a test-taker has to appear for an exam for some formal reason (sometimes the case at universities), it seems fairly unlikely in small low-stakes contexts.

*"Baxter ... about this report ... your punctuation, spelling and grammar are perfect. No one can understand it!"*

Decide what each score means. Start with the highest and define this, then define the lowest score. Define the other scores and then decide if you need more stages or if you have too many. Change the number of stages for each criterion if necessary. Do not leave scores undefined or allow half points without identifying them. All raters have a tendency not to want to come down on one side or another, just as they are often reluctant to give full points. If what has been written is worth full points then it should get these.

**To think about**

If you have half points you will have to decide what to do with half points at the end – do you round a half point up or down? What implications does this have for the test-taker?

Although criteria are the most important and most obvious way of ensuring objectivity in tests of production, another way is to also make sure that trainers do not mark their own learners' tests. If it is not possible to find someone else to mark them, then one way of making the marking more objective is to ask the learners to put numbers rather than their names on the papers. Ideally, all tests of writing should be marked by two or more raters independently of each other. None of the raters should know or be able to identify the test-takers. This may seem idealistic, but one of the following should be possible in most BE testing situations. In all cases, no raters or as few as possible should know the test-takers.

- Two raters with a third to decide if the two scores are not the same
- Two raters who are obliged to agree on a final score
- Two raters whose scores are averaged
- Two raters, where one has the final say on the score
- A group marking session where questions and anything which is unclear can be dealt with immediately

Raters will also need help and instructions for marking, even if they have attended a benchmarking or rater training session. Put the criteria in a logical order for marking, i. e. comprehensibility of the main message before details of grammar or spelling. Make sure that each rater is provided with an example of the task and any details on marking which are not included in the criteria. It is often useful to provide outline criteria which are used for actual marking, with notes and examples for each one, to refer to if necessary.

Remember that the more detailed the input, the more uniform the responses from test-takers will be. This will make this type of task easier and quicker to mark. If essay questions are used (these apply more to university BE tests than in-company tests) where only a title is given, what test-takers write will depend on their age, background and culture as well as their writing skills, and additional aspects of writing will have to be taken into account. Each task will have to be considered independently, and if two tasks are used in one test which are different from each other (which ideally they should be), different criteria will have to be developed for each.

## 6 Accuracy

How accurate should the piece of writing be (how many mistakes are acceptable and how do we score them)? It is probably safe to say that most English teachers, and BE trainers are probably no exception, come down strongly on linguistic mistakes in tests. Do not let your trainer's view influence the score you give for each criterion. The writer may get the message across appropriately in spite of mistakes. These mistakes may not be noticed by readers, not even if they are native speakers of English.

When defining the criteria for accuracy of language, it is important to think of what you mean by mistakes and their number. The phrases *a number of mistakes* or *several mistakes* are often used in criteria as well as the phrases *serious mistakes* and *basic mistakes*. Although these are useful as short definitions, they should be more closely defined somewhere in the marking instructions or in rater training. The number of mistakes can be defined in relation to the total number of words or, in a short piece of writing, as an absolute number. How basic the mistakes are will depend on the level of the test-takers, and how serious the mistakes are will depend on a definition. This will have to be provided. Perhaps in view of the increasing emphasis on communication in language teaching, this definition should not be based only on grammatical progression and teachers'/trainers' notion of what a serious mistake is.

> "One way of defining how serious a mistakes is, is to define the extent it confuses, amuses, irritates, offends, annoys or misleads the person who hears or reads it."
>
> Adapted from the CEFR description of sociolinguistic appropriateness B2

If you or other trainers are tempted to modify the criteria to suit your own ideas, remember that tests marked in different ways mean that

- all the results are then unreliable or objectively meaningless,
- the test-takers' time has been wasted,
- all the raters have wasted their time marking the test, even those who have used the criteria as they were intended,
- test-takers can legitimately complain about their results, especially if they compare them.

Rater reliability is one thing that non-experts in testing can understand.

## 7 What are the advantages and weaknesses of tests of writing?

The advantages or strengths of tests of writing are that the skills can be tested directly and the tests can be made very authentic. Also test-takers cannot guess in tests of writing. It is fairly easy to produce tasks as well as model answers. Test-takers are actively involved in the production of language and are given an opportunity to show what they are able to do.

Of course tests of writing penalise test-takers who do not write well in their own language and tend to favour those test-takers with strong analytical and reasoning skills as well as good knowledge of grammar and vocabulary. This may be the case but if the outcome of a task is the most important indicator of success (competence in the language in this case), then setting tests of writing is completely justifiable. If a person needs to be able to write a report for a certain type of job, it needs to be tested, no matter what their background is.

■ The work involved in setting tests of writing is mostly to do with the criteria to ensure that marking is not and does not become subjective. One feature of tests of writing (which may be considered as an advantage or as a disadvantage depending on how you look at it) is that test-takers have more control over what they produce and can write what they want (up to a point). On the other hand, the range of topics given to test-takers in a test of writing can be restricted by the test setter. Perhaps the most important thing to remember is that although tests of writing can provide information about test-takers' grammatical knowledge and vocabulary range, as well as other things, such as power of expression, they can not be used to make assumptions about oral skills. ■

### To think about:

- What are the main writing tasks your learners will have to do?
- What criteria do you think are important for writing?
- How important is accuracy?

CD-ROM

*"… oral tests, because they involve a subjective judgement by one person of another, are likely to be less reliable; … the human aspect of that judgement is precisely what makes them valuable and desirable."*

Underhill, Nic (1987): Testing Spoken Language. CUP, p. 5

# 10 Speaking

## Testing speaking is a unique opportunity

- Why is assessing speaking a challenge?
- What is particular about speaking?
- What should the examiner's role be?

Whether speaking is tested at all may depend on the view of language, the educational culture or the genre involved. In some countries the main focus in language teaching is on writing and reading, and in some genres, for instance academic English, speaking does not play an important part. Tests of speaking offer unique opportunities for learners and testers, as there are several things which cannot be observed or tested in other ways. For this reason it makes sense to concentrate on these aspects of language in tests of speaking, both in the tasks and the criteria. Tests of speaking also usually entail interaction, whether between examiners and test-takers or between test-takers. It is the different facets of tests of speaking which make them particularly interesting and challenging to learners, teachers and testers. This chapter will look at testing and assessing speaking as an isolated skill and consider some of these facets.

There may be good reasons for not testing speaking. However, it is often neglected in informal or low-stakes language tests, even when it is the most important skill taught and practised in classes. Many informal (and some formal and large-scale) and many placement tests do not have a speaking component, even when speaking is to play an important role in teaching. Where assessment of speaking does take place, often little attention is paid to the really important aspects of this particular skill, whereas too much emphasis is placed on elements which can be better assessed elsewhere. There may be

reasons, although not necessarily justifications, for not testing speaking explicitly. Although tests exist which claim to provide a valid and reliable assessment of speaking skills by testing these indirectly, i. e. not by asking test-takers to actually speak, these are not the focus here. The reason for this is that

> " … to determine whether learners can communicate orally, it is necessary to get them to take part in direct spoken language activities."
>
> *Weir, Cyril (2005): Language Testing and Validation. Palgrave*

There are sometimes purely practical reasons for not testing speaking as often and as much as other skills. Questions of organisation and logistics have to be dealt with. These as well as other things make testing speaking different to testing other skills. On the other hand, testing speaking also offers teachers, testers and learners/test-takers several new opportunities and chances which cannot be obtained by using tests involving only reading and writing.

## 1 What opportunities do tests of speaking offer?

When speaking is assessed or tested, it is still often graded according to the grammatical correctness of what is said. Many teachers and testers expect full sentences to be produced in tests of speaking, and this with minimum hesitation and using correct language. The test-taker is expected to understand and adjust to a task, often on the spur of the moment, formulate his/her ideas and then produce them in correct language while paying attention to pronunciation and intonation at the same time. Looking at what is expected of test-takers in many tests of speaking often reveals how little these tests fulfil two of the criteria for good tests: authenticity and interactivity.

▶ Chapter 1

Similarly, looking at what is unique about speaking and how this relates to testing can show what the opportunities for tests of speaking are and how little these are taken advantage of. For this reason, the observation and testing of grammatical correctness in tests of speaking play a relatively small role in this chapter.

▶ Chapter 12

## 2 What is unique to speaking?

### Pronunciation

The first thing is that speaking involves using a sound system and getting your mouth around words and sounds. This in itself is not easy and requires practice, without in addition having to consider the content of what you are saying. In general, as well as existing standards of pronunciation, there is a very wide range of pronunciations which can be understood by most people.

If the reason for speaking is to get a message across, then pronunciation need only be considered in connection with how it influences this message. Views on this vary considerably and will influence how you choose to consider and mark pronunciation. You may wish for a near-native pronunciation and intonation, you may be content with simply comprehensible pronunciation.

In a test, all those involved will have to agree on what sort of pronunciation they want and can expect. Even the term *near-native* is not easy to define as it depends on what sort

of native speaker (GB, US, Australian, South African etc.) is meant as well as on how near the pronunciation should be to an agreed standard. Particularly with intonation, comparison with native-speakers may not be relevant. Alternatively, the pronunciation and stress of individual words may be more important than overall accent or intonation.

**Remember**

Examiners are expected to keep to the instructions for the tasks and mark according to the criteria in tests of speaking. This is essential, whether or not they have set these themselves, to ensure validity, reliability and fairness.

Pronunciation is something which obviously applies only to speaking. Clearly, if the examiner cannot understand anything the test-taker says, no other criteria can be marked. This is a fairly clear-cut case, which however will not probably apply to BE (assuming it is being tested at a fairly advanced level). At high levels, examiners tend to mark test-takers down for mispronunciations of individual sounds, letters and words as well as for mistakes in intonation. Assuming that the test of speaking has not been recorded, marking pronunciation often leads to discussion among examiners: usually no one is absolutely certain afterwards of what was actually said during the test. Even two examiners cannot always agree whether a particular word was pronounced correctly or not. On the other hand, a test of speaking is the only place where pronunciation can be marked.

**To think about**

Excessive emphasis is sometimes placed on certain sounds (such as 'th'), often to the exclusion of other, equally if not more important, aspects of pronunciation.

Certain sounds are in themselves difficult to make (even for English children when they learn to speak, and are often "mispronounced" by some native speakers throughout their lives). Also achieving native speaker-like pronunciation may not even be an objective for all foreign language learners. Pronunciation and accent modification is often a question of talent and/or excessive concentration on this aspect of language. It is true that sounding like a native speaker can compensate for other mistakes and generally makes a good impression, but it should also be considered whether sounding like a German, Frenchman, Japanese or whatever is a good reason for penalising a test-taker.

One thing that can lead to confusion or misunderstanding is wrong syllable stress in words. This is perhaps something which learners can be expected to learn and assimilate, especially with the possibilities available of listening to words in electronic dictionaries. This is connected to learning vocabulary. Learning a new word should include learning how to pronounce it, especially if it is to be used in spoken language.

### Speaking is subjective

Another challenge in testing speaking is that the marking is often considered highly subjective. It is subjective in that the speaker chooses what to say from the repertoire he/she has available. The reasons for making these choices are subjective; they may depend only on what the speaker knows, but may also depend on other factors such as the setting, the content and the person being addressed. Each of these may also be affected by the way the speakers see them. They may see a setting as formal or informal and as requiring appropriate language for this, they may feel they do not know the right words for the content of what they want to say. They may assume the other speaker feels or does not feel the same way as they do. The person or people being addressed will have their own (possibly highly subjective) views on the matter, and all these will influence not only what is said and how it is said, but also the effectiveness of the communication. Finally, whether all the participants see the communication as effective or not will also, in the end, depend on each individual viewpoint!

However it is certainly not the case that all assessment of speaking can only be subjective. This is demonstrated by all reputable exam boards that produce well-thought out speaking tasks, exam procedure and criteria for marking. Producing these, in particular the criteria, takes time. It requires that all those involved stick to the task, procedure and criteria agreed on to ensure that the test fulfils the criteria for a good test.

▶ Chapter 1

In many informal or low-stakes tests of speaking, it is only the examiners' reactions which play a part in awarding marks and grades. These reactions are often not based on real-life reactions, but are generally the reactions of language teachers to correct or incorrect grammar, pronunciation and vocabulary. The examiner's role plays a far more important part in tests of speaking than in any other tests.

At one time the term *naïve native speaker* was commonly used. This does not mean anything about the native speaker's intelligence or common sense, but was used to describe a native speaker who had no knowledge of the test-taker's native language and no particular knowledge of language structure or language learning. It is usually fairly easy to imagine someone like this, a member of one's family or friends in Britain or America, and to imagine what their reaction would be to what is said. Would they understand it? How well? How would they react?

The subjective nature of speaking means that setting and keeping to standards in procedure, marking and grading are the only way to ensure the highest degree of objectivity and fairness to the test-takers. The setting of standards is not a democratic process. In the end someone will have to make the decision on which standards to use. This may be the individual BE trainer for tests in one group or a head of department or language coordinator for all learners in a company or university. Standards can be considered good or bad by individuals, but this does not mean that they do not have to stick to them during the test. If a test of speaking is being carried out with a number of examiners and test-takers, it would be unfair if the examiners were treating, assessing and marking the test-takers in different ways.

▶ Chapter 14

## Speaking is transient

The nature of most speaking is that it is spontaneous and transient. This is not always the case, but in a wide range of speaking tasks it will be so. Even when what is said has been learnt by heart or is being read out, once the words have been spoken, they cannot be changed. A mistake can be corrected, but it has still been made. This can be a disadvantage but also an advantage. Parts of what has been said will be forgotten or ignored. Unless what is said is recorded, and this is still rarely the case in tests of speaking, there will probably be no way of reconstructing the exact words spoken. What will usually be remembered most clearly is the message. The message consists basically of the sounds (pronunciation) and the content, and the construction of this content by the listener. This may involve the vocabulary and grammar. The interaction between these two is extremely important. If I cannot understand what someone is saying because of his/her pronunciation, the message will not be conveyed. However, even if the pronunciation is completely understandable, this is no guarantee that the message will reach its recipient.

**Did you know?**

The Council of Europe produces and distributes DVDs with recordings of candidates at the different CEFR levels which demonstrate this point quite clearly.

In some ways this is similar to aspects of writing in that illegible handwriting and incomprehensible or misleading spelling and punctuation will prevent the message from being understood, but formally correct writing is no guarantee that the reader will know what is meant. In both writing and speaking, effective communication depends on speaker *and* audience.

## Reacting

Speaking can consist of simply producing spoken language, but in BE and at fairly advanced levels, it also means reacting. Although a test of speaking can consist, for instance, of reading a text aloud, this not usually considered spoken interaction. In most cases, speaking means reacting fairly quickly to input, whether spoken, observed or

implied. Everyone has certain expectations of the time required before a reaction takes place. These expectations depend on the particular situation and on the speakers and listeners involved and will vary from culture to culture, but always mean that a speaker has to have language at his or her immediate disposal, even if it is only to maintain the communication or explain that he/she doesn't know the word to use or hasn't understood something. At low levels, the language may consist to some extent of body language, but in BE communication, effectiveness will definitely be impaired if speakers' reactions are considered too slow. There is less time to think in spoken communication than in most written communication or when reading, which makes testing speaking in some ways similar to testing listening.

▶ Chapter 8

Spoken interaction usually (if not always) involves listening and decoding. However it also simultaneously involves formulating and reacting.

Reacting in a timely fashion means not only knowing how to say something, but also what to say. Although this is not a linguistic issue, it often plays a part in tests of speaking. A reaction may be delayed for several reasons. The speaker

- does not want to speak,
- has not understood what he/she has heard,
- has not understood that he/she has to speak at this point,
- does not have anything to say on the matter or
- does not know how to express it appropriately.

In a foreign language, the situation is made more complicated by the interplay between the reasons. Foreign language users may not have understood something but cannot express this appropriately in the foreign language. Equally, they may know exactly what they want to say, but not how to say it in the foreign language.

### Interaction

Speaking involves others. It is usually the person speaking who is the focus of tests of speaking. However, as well as the speaker, the others involved in the interaction must be taken into consideration. Effective spoken communication is always two- or more-way. Its success depends not only on the speaker but also on the person(s) at the other end of the communication, whether one, two or hundreds. They may speak less than the (main) speaker, they may speak very little or even not at all, but it is rare (except perhaps in tests of speaking!) that there is no reaction at all to what is said.

There are cases of this, for instance when listening to a lecture or a radio programme, but even members of an audience react in some way, perhaps by speaking to each other or by means of non-verbal reactions or even with boos and catcalls. In some cases the reaction to what is said may only be an imagined or predicted one (for instance when recording a message for an answering machine or making a video recording for Youtube or for a job application), but in general the reaction is felt fairly immediately by the speaker.

This reaction to what is said may not be expressed in words. However causing some sort of reaction is usually one of the main reasons for speaking. Achieving a reaction is, similarly, not a one-way process but the further course of the interaction. The reaction a speaker receives can affect all aspects of what he/she goes on to say, including pronunciation, fluency and grammar. And yet, many tests of speaking fail to take the real-life reaction of a listener or the effect this has on what is said next into account. It is this co-construction of meaning, the collaborative or cooperative nature of speaking, which makes testing speaking different from testing reading

or listening and even different from testing writing, despite the similarities between the two as productive skills.

All these aspects of speaking apply equally to spoken communication between native speakers of the same language. The fact that speaking is a complex process clearly does not mean that communication is impossible. Spoken communication, in whatever language, goes on all the time, which is why it is generally considered to be the most important part of language and language learning, but clearly, it is not easier when a foreign language is being used.

**Remember**

The role and behaviour of the examiner is extremely important. If the examiner's role is only to assess, evaluative remarks (*Good, yes, fine*, etc.) may well affect what the test-taker says next. It is highly unusual for the interlocutor in spoken interaction to be simultaneously assessing what the partner is saying.

## 3 Tasks for tests of speaking

It is easiest to decide on the tasks first when designing a test of speaking and then to proceed to the criteria, as these will depend partly on the tasks set. In general, teachers find it fairly easy to think of authentic and interactive (in the testing sense) tasks for tests of speaking. Many tasks in tests are much the same as tasks used in class. Which ones you choose will probably depend on administrative and practical issues. The main thing about tasks to be used in tests of speaking is that they are valid and authentic, i.e. they should reflect the sort of speaking which test-takers will have to do in real life.


*The tasks should reflect the speaking the test-takers have to do in real life.*

**Remember**

Validity also applies to the marking criteria!

Some examples of exam tasks are:

- An interview on a subject with the examiner as interlocutor
- A discussion with the examiner(s)
- Giving appropriate responses to situations provided by examiner(s)
- A presentation (of a company, a product) with examiner(s) as audience
- A discussion with one or two other test-takers on topics provided by examiner
- A consensus-finding exercise with one or two other test-taker(s)
- A group negotiation
- A group role-play

Tasks which can be used in BE tests can be categorised under the following headings. BE trainers will probably have examples of all these different tasks in their course material:

- Presentation/monologue
- Negotiation
- Collaboration/cooperation
- Discussion
- Transaction

The decision as to which task(s) to use will depend on various things such as the expected outcome, the examiners and other practicalities.

It may be helpful to start with the outcomes test-takers want. What the learners should be able to do orally in real life should form the basis for the decision on what task(s) to use in the test. Test tasks should replicate the real-life situations of the learners as far as

possible, which, as already mentioned, makes finding tasks for tests of speaking relatively easy. What remains is to decide how the tests can be conducted and marked so that they will be reliable and fair to as many test-takers as possible.

**To think about**

If the examiner is the interlocutor, he/she may also have to play a role. Is this realistic and is the examiner able to play the role and assess the test-taker's performance properly?

The roles the test-takers play in a test of speaking are usually fairly straightforward. They perform a task using speaking skills to do this. The roles they take on are connected with this task and they may be themselves or, in a role-play, play a different role. Here again, it is important to remember what is being assessed. Test-takers should not be asked to play roles which are very far removed from their real experience. Even if this is done in class, a test is not usually an appropriate place to demand imaginative or creative as well as acting skills on top of those involved in producing language.

Reading aloud as a test of speaking is worth mentioning here as it is still practised in class and often used as a test task, even though there are relatively few situations (especially in BE) where reading a text aloud is appropriate. Even if reading aloud is used to mark pronunciation, the pronunciation of natural speech should also be marked, as often the two vary quite considerably. The time used to assess texts read aloud could possibly be better used for a more communicative and authentic task. The marking is likely to be just as reliable and the test probably more valid.

It is sometimes felt that speaking can be marked just as well in a course during teaching time and does not need to be tested explicitly. There are, however, good reasons for setting aside a task, a time or a specific setting for a test of speaking. Learners in general perform better when they know they are being evaluated, i. e. something depends on the outcome of the task. In this way, they are simulating a real-life task, where certain rules have to be followed and their performance and the outcome (e. g. of a business presentation or a negotiation) are important.

## 4 Marking a test of speaking

**Remember**

Different criteria for good tests affect each other.

▶ Chapter 1

Testing speaking is not the same as just letting test-takers speak and seeing how they do. A test is always a compromise governed by rules and it is the setting of the rules which is important if the test is to be both authentic and valid. In tests of speaking the rules may make the test less authentic and interactive but more valid.

Assessment of speaking is called different things by different testing bodies, schools and teachers. Some of the names used refer to the formal function (*Speaking Test*, *Oral Exam*, *Test of Spoken English*), some to what happens in the assessment (*Interview*, *Role-Play*, *Discussion*, *Presentation*). In all cases, test-takers are expected to produce spoken language, and this is assessed by examiners. This is what is meant by a *test of speaking* here. So that speaking is not marked entirely subjectively, it is important that criteria are developed which examiners, whoever they may be, can use. The two issues, of how the tests is marked and who marks it, must be considered separately to ensure that all test-takers are treated fairly.

In tests where speaking is being tested directly, the main thing is that test-takers speak. If no spoken language is produced, the test cannot be marked. Speaking is often the skill that learners find most difficult, some are even reluctant to do it until they have mastered other parts of the languages and other skills. This reluctance often comes from the fact that their spoken performances, for instance at school, have been assessed and graded using criteria for other skills, such as writing or reading or on their use of correct grammar. Marking criteria are particularly important in tests of speaking. Deciding on

the criteria means making a decision as to which of the variables of speaking are to be tested and marked.

▶ Chapter 13

Pronunciation, task management, interaction and fluency can be best (or perhaps only) assessed in tests of speaking, which makes it strange that often little attention is paid to them in classroom tests. In many tests of speaking, linguistic accuracy is the focus of marking. Counting mistakes in grammar and vocabulary can lead to fairly reliable and objective marking if all examiners follow the same rules for scoring, but it will not make a test of speaking valid, authentic or interactive. It will only produce comparable results for test-takers on one aspect of their performance which has little to do with real-life communication. It will lead to a test which is not authentic because the test-takers will not concentrate on what they say but only on how they say it, and the test will not be interactive because it is asking test-takers to perform in a way they would never have to in real life.

**Example**

Basing the marking of a speaking test entirely on linguistic accuracy is equivalent to giving the following instructions:
*I will ask you some questions about yourself. Make sure you answer in complete sentences and do not make any grammatical mistakes.*
Or
*Discuss the situation described here with your partner. Ask your partner for his/her opinions and give your opinions. Make sure you only use correct grammatical structures and pronounce all the words correctly.*

## Pronunciation

In general the criterion for pronunciation should set out how far the following, among other things, lead to a breakdown in communication and how much and how they should be penalised:

- Mispronunciation of individual sounds, e. g. 'th', 'w'
- Wrong syllable stress in words, e. g. *executive, determine, hotel*
- Accent in general

Taking all this into consideration means that when developing marking criteria for a test of speaking, pronunciation should (probably) be marked but

- should not be weighted too heavily, i. e. carry too many points,
- test-takers can be expected at a certain level to know where the syllable stress is on words relating to their area of work, and
- test-takers should be given the benefit of the doubt, especially when they only mispronounce something once.

## Task management

When grading speaking, interaction is sometimes reflected in what is often known as task management, i. e. the successful achievement of the task. Depending on the particular task, the test setter will have to consider what contributes to its achievement. Is it the use of polite language strategies, for instance, or is it the length of the test-taker's contribution? Depending on the task, the speaker may need to be short and to the point or may have to provide a long and detailed piece of spoken language. Can the message be understood immediately? Is a checking question necessary?

Giving scores for interaction to individual test-takers is the examiner's main task and requires some concentration. Although criteria and examiners' use of them can aim to be objective, the interactive skills of any language user will always depend on other factors, some beyond our control as trainers, test setters or examiners. These include personality, mood, the other participants in the interaction, strategic skills and identification with the task. Again, however, tests of speaking are the only place where interactive skills can be assessed, and so some inconsistency and subjectivity may have to be tolerated.

Depending on the task, interaction and task management may also depend on contributions of others such as the examiner(s) or other test-takers. This must be taken into account when designing the task, defining the examiners' or other test-takers' roles and organising the test. In general, test-takers are marked individually and although there are schemes for group marking, this is a rather complicated and fraught procedure, especially if there are high-stakes involved. Awarding one mark jointly to a group of test-takers is generally not advisable.

## Fluency

An essential factor in effective spoken communication, as mentioned above, is reacting quickly, obviously within certain limits and depending on the situation. An appropriate reaction is timely and expressed in such a way that the entire message can be grasped at once. This is what is meant by fluency. In general, people are regarded as being able to speak a language fluently if they can react quickly and do not have to search for the right language to use. Language teachers often mean that they do not make grammatical mistakes, but this is not a criterion that other people often apply.

A mark given for fluency should take into account the nature of the task as well as the level of the test. In BE, speed of reaction is more or less essential, whether it is to keep a social conversation going at a business dinner, or to react to an order or a complaint. As well as this, at higher levels, test-takers should be able to fill gaps with language using strategies to gain time, using expressions such as "Let me think ..." or "That's a bit difficult ...".

## Grammatical accuracy

Does this mean that grammatical accuracy has no place in the marking criteria used for a test of speaking? Most tests of speaking do have at least one criterion which is related to linguistic accuracy. However, as we have seen, pronunciation, fluency, task management and interaction are criteria which apply almost exclusively to speaking, so that assessing grammatical accuracy or range of vocabulary should not take place at the expense of one of these. In a test where marking is taking place in real time and there is no way of repeating the test, examiners should not have to deal with more than six different marking criteria at the same time. Four or five are better. This means that it is definitely worth thinking carefully about how far grammatical accuracy should be a criterion for a test of speaking, particularly if it can be tested elsewhere. If you do include grammatical accuracy as a criterion, ask the following questions when deciding on weighting and scoring:

- Is the number of mistakes important? You may want to take the level of the test-takers into account and decide what sort of mistakes they should no longer be making, or you may just want to consider the number of mistakes made. In the case of BE at a high level, you may also want to consider the impression a lot of grammatical mistakes may make on a business counterpart (although this is probably not as serious as trainers think).

- Could the mistakes lead to misunderstandings? In fact there are relatively few mistakes which lead to real misunderstandings. These may be more important in terms of interaction and task management.
- Does the test-taker take risks, i. e. use complex structures to convey complex content, and make mistakes because of this? If this is the case, then again, the questions above should be answered.
- Is the speaker very fluent? In general fluent speakers will make more mistakes than those who are very hesitant or speak slowly and think a lot before they speak.

Criteria can influence each other. For instance, a test-taker may be very fluent, but make many grammatical mistakes. Another may make no grammatical mistakes but not contribute to a successful task outcome and interact inappropriately (e. g. be impolite) while remaining grammatically correct. When deciding on your criteria, make sure that you are marking valid aspects of speaking (in general) and the task (in particular) and that you are not marking the same thing twice. Even if the criteria are for two different aspects, examiners should be careful not to reward or penalise the same behaviour twice.

The aim of criteria in speaking, as in writing, is to ensure validity (that the right things are being marked), reliability (that all examiners are marking in the same way), and as much objectivity as possible.

▶ Chapter 13

## 5 Test-takers

One major decision is how many people are to be tested at the same time. If a test of speaking is to be authentic, it may have to involve at least two people. This will be the case when negotiating skills or the ability to take part in a meeting are being tested. Tests of this type can involve a group of test-takers interacting with each other. Making this as authentic as possible will mean arranging the furniture in as authentic a way as possible and making sure that the test-takers are not (too) aware of the examiner(s) marking their performance. It is quite possible for one examiner to assess and mark two or three test-takers if the marking criteria and the examiner's role are clear. With more than three test-takers, it is advisable to have two examiners, although this will also depend on the length of the task and test. It is possible for one examiner to mark up to nine test-takers if they are all interacting, the examiner is only assessing and the test task takes at least 30 minutes to complete.

The most valid, reliable, authentic and interactive test of speaking which is also fairly practicable is therefore one in which two or three test-takers perform a communicative task with one or two examiners marking it. The advantages of this format are that it represents a fairly realistic situation. Test-takers can interact with each other fairly naturally without being too conscious of being assessed. In real life, the person speaking to you does not usually assess your language at the same time as the communication is taking place, and an examiner making notes while speaking will place some strain on the interaction. Having the examiner play the role of interlocutor also tends to make the roles in the conversation less natural as the test-taker tends to defer to the examiner and is less likely to initiate, ask questions, criticise or disagree.

A test of speaking, however much you attempt to make it as authentic as possible, will always be a performance, just as a driving test is. This has the advantage that test-takers try their best to show what they are able to do, possibly using language and gambits they

would not always use in real life. It can however have the disadvantage that the test-takers are nervous and make more mistakes (of whatever kind) than they do in real-life situations.

Most situations where learners of English have to speak are performances of one sort or another. There is nothing inherently wrong with test-takers learning language chunks, gambits, or even stretches of languages off by heart for a test. This is in fact what we do all the time in real life, especially, for example, when giving a presentation. If the language produced seems to be natural, then it does not matter how it has been learnt, as it is the performance and not the learning which is being tested. The speakers can still be assessed on their ability to deal with unexpected situations which arise. How spontaneous speech comes about in a test will depend on the nature of the task, the roles of the other test-takers and the roles the examiners play. One aim of a test of speaking is to allow test-takers to produce both rehearsed and spontaneous speech in as natural a setting as possible.

Just as natural communication sometimes breaks down, communication in a test may also break down. One reason for using examiners as interlocutors is that they can deal better with a breakdown in performance, but it is worth considering, especially in BE at a high level, what a breakdown in performance consists of and whether expecting test-takers to deal with difficulties could not be part of the assessment. Many situations in business involve dealing with difficulties and misunderstandings, and one major criticism of many tests is that they do not deal with problems realistically.

Designing a test of speaking, where speaking is understood as oral performance, is as easy as designing any other test. The focus is just slightly different. Tasks for tests of speaking can be found in real life and in course books. What require more thought are the role of the examiner(s) and the conduct of the test, as well as the criteria to be used to mark the performance. All aspects of the test, criteria, task, timing, and the roles of those involved need to fit together to make the test a good one. The criteria have to be decided on, bearing in mind that not all aspects of speaking can be realistically marked simultaneously in a short space of time, and that some aspects of language use may be better marked in tests of other skills, not because they are not important in speaking, but for practical reasons. The main criteria used in marking tests of speaking are usually taken from the following: fluency, task management, interaction, pronunciation, range of active vocabulary and grammatical accuracy.

Making a test of speaking valid and reliable and as objective as possible while remaining practicable to conduct may be a challenging task but not insurmountable and certainly worth embarking on in view of the role that speaking plays in the general view of what using language means.

## 6 Examiners

**What is it?**

The word *examiner* is used here throughout for the person conducting and/or marking a test of speaking.

In the same way that tests of speaking have different names, examiners are also referred to by different names such as *assessor*, *interlocutor*, *interviewer*, *observer*, *scorer* or *rater*. Usually the names refer to the function. In most cases the examiners have more than one function in a test of speaking, the two most common being *assessor/scorer/rater* and *interlocutor*. Several functions may be performed by one examiner, or they may be divided up between two or more examiners.

The interlocutor is the examiner who interacts with the test-taker and plays a part in the spoken communication, rather than simply observing and/or marking. The interlocutor's role may be that of interviewer or of conversation partner, and what the interlocutor is expected to say may be scripted in entirety and binding. In other cases, there may be notes for the examiner to use in formulating what he/she says, or this may even be completely left up to the individual examiner. All these things will influence the course of the test.

Testing speaking is much easier if only the effectiveness of the message is assessed. However, it is probably not possible to find enough people to act as examiners who agree not only on what "effective" means in general, but also what it means in the particular case concerned. Examiners could simply be allowed to mark subjectively, according to their own personal views. This is often what happens, but is not only not valid, reliable or objective, but also unfair if different examiners mark different test-takers. For instance, a test of speaking could be that each examiner just talked to test-takers about anything they wanted, in order to see how the test-taker managed (i. e. how well they were able to react spontaneously), but this would probably not always be valid and would certainly not be reliable as the interaction would take a different course with each test. Although it might be realistic, all the variables involved would have to be taken into account in the test format and marking instructions.

As there are no right answers to agree on and look for in tests of speaking in order that test-takers are treated fairly and get the results they deserve, examiners need particular qualifications and a level of awareness about their own, probably subjective, views of the task, procedure and performance of the test-takers. These requirements for examiners can be summed up as follows:

- Examiners must be aware that any judgements of spoken language are likely to be highly subjective.
- Examiners must be willing to step outside the role of language trainer and imagine themselves in the role of recipient of the message.
- Examiners must be willing and able to stick to standards, even if they do not agree entirely with them.

The main practical issue is probably how many examiners can be involved. As in all tests, the more examiners marking the performance, the more reliable and objective the result will be, but the longer it will probably take to reach a consensus between examiners on the marks to be given. If two examiners can be involved, it may be helpful to give them different roles during the exam, but this again will also depend on the task and the number of test-takers being tested at once. There is also no reason why the examiners' task (however many examiners there are) should not be restricted to assessing, i. e. their main role. There is no real need for examiners to take part in assessment tasks, and it is only in tests of speaking that this happens at all. Examiners' influence on the outcome of a spoken test can be positive or negative, but there can be no doubt that the influence of a teacher, a competent speaker and an assessor will alter the nature of the interaction and probably make it less authentic. There are few situations outside educational settings where one of the performers also assesses the performance!

The advantages of the examiner playing an active part in the interaction are that he/she can keep the situation under control, so there is less danger of communication breaking down or going in an unintended direction. Another argument for using an examiner as an interlocutor is that the test-taker has a reliable model to copy, if necessary, and will not adopt possible mistakes (usually only grammar or pronunciation mistakes are meant

here) made by another learner. Further arguments against letting learners interact with each other in a test are that big differences in levels can have a negative effect on potentially stronger performances (although it is also said that weaker learners can benefit from stronger ones) and that personal antagonisms or differences in hierarchy may affect performance. This, however, can be avoided by asking test-takers to choose their own partners. Doing this is also one way of making a test a less nerve-racking experience and giving test-takes the opportunity to perform to their best advantage.

**To think about**

Assessment centres are part of many business people's lives nowadays and often used in job application and selection procedures. Tasks from these are used in some course books and can often be adapted to BE testing.

The roles of the examiners can range from only conversational partner to only assessor, with others such as interviewer, interlocutor, counterpart, moderator and guide and any combination of these in between. Any part the examiner plays should not affect the main point of the test, which is the assessment of performance, so deciding on the role of the examiner must take this into account. Most BE examiners find it easiest to assess if they do not have to take part in the interaction and can just observe it. How easy this is will depend on the task and the number of test-takers being assessed at once.

Once a decision has been made, especially if several groups are being tested by different examiners, then all the examiners should stick to the same rules of behaviour. Many examiners are not aware of how much their behaviour can affect the test-takers' performance and thus their marks. Here are some examples:

- In a pair exam, if the examiner joins in, however well-intentioned this may be, the test-takers will immediately stop talking to each other and start talking to the examiner.
- If the examiner frowns, the test-takers will assume it is because of their performance.
- If the examiner says "Okay, good, let's go on to the next part", the test-takers will be surprised when they get a low score.
- If what the examiner says is not scripted, some examiners will use simple language to explain things and others complex. Even if the examiner's role is scripted, some will speak slowly and others quickly.
- The examiner may have to establish a relationship with the test-taker before going on to the task. Some examiners will be able to do this better than others.

**Remember**

Be friendly! Just because you are marking a performance, there is no need to be unfriendly or strict in your behaviour. Making test-takers feel at ease will help them to do their best.

There are many more such examples. Not all of them can be avoided or prevented by the test setter. As mentioned above, tests of speaking will always retain a subjective element, so it is important that examiners are aware of this and that it is kept to a minimum. This is one instance where examining is quite different from teaching. Examiners in oral exams must be prepared to stick to rules of behaviour (these may even be self-imposed) to ensure the validity of the test. This is why reputable exam boards insist on examiner training and carry out inspections, sometimes even regulating what words examiners may and may not use. In an individual test of speaking, making the examiner use a script may be inauthentic, but it goes some way to making the test objective and reliable as well as fair to all the test-takers.

From a test of speaking
Examiner: "I'd like you to tell me what you would say in the following situation.
A friend is visiting you and you think he may be thirsty. What do you ask him?"
1st Candidate: "How old are you?"
2nd Candidate: "Do you want a bath?"

## What qualifications do examiners need?

Most examiners are teachers, so are qualified in this way, i. e. they are competent speakers of English and know about the area of language they are examining. However they should

also be able to take one step back from their roles as teachers and observe the instructions and criteria they have been given as examiners. Depending on what these are, it may mean, for example, that they should not help the test-takers with words they do not know and/or should not give any feedback on the results. In order to be able to follow the rules, examiners should actually be familiar with the construct of the test and the rationale behind it. Whether they agree with these or not is secondary, as, most importantly, with several examiners and large numbers of test-takers, the results must be reliable and fair.

Whether examiners for tests of speaking have to go through training and how extensive this training is will depend on various factors. It is always important to choose competent examiners. This will always involve being familiar with the test tasks, the level of the test-takers, the role of the examiner(s) and the criteria.

Differences between different examiners' perceptions of these should be dealt with before the test and standardisation achieved as far as possible. Ideally benchmarking sessions should take place between tests to maintain the standards set.

■ All of these considerations will be of some importance in tests of speaking, although the degree to which decisions have to be made on them may differ from case to case. However, even when simply making an informal or individual judgement on how well someone can speak English, it will be necessary to think about what aspects of speaking are being considered or judged, what role the setting plays and how much any other people involved influence the language produced by the speaker. It is these considerations, formalised, which make up the basis of a test of speaking. ■

### To think about:

- How can a test of speaking for BE be valid, reliable, authentic, interactive and practicable?
- Do you want to assess whether test-takers can perform a task or whether they can produce correct language?
- What requirements do you have for examiners?

CD-ROM

*"Vocabulary is now considered integral to just about every aspect of language knowledge."*

Daller, Helmut / Milton, James / Treffers-Daller, Jeanine (2007): Modelling and Assessing Vocabulary Knowledge. CUP, p. 1

# 11 Vocabulary

## Language is more than just words

- Can words be tested authentically in isolation?
- Why do you want specifically to test vocabulary?
- What is the best way of doing this?

Although grammar was, has been and still often is the focus of language teaching, it is evident to many that vocabulary is more important in producing language for communication. Asking for something without knowing the word for it is almost impossible even if I know the grammatical structure necessary for making the request. In particular in BE contexts, it is impossible to function without the appropriate vocabulary. It may be a case of recognising the vocabulary, identifying the correct meaning from a range of possibilities, or it may mean using the right word at the right time and in the right place. This may be a question of meaning, but also one of register or usage.

**Example**

The difference between *stay* and *remain* is not one of meaning but register or collocation.

What in particular is to be tested will determine where vocabulary in BE is tested, whether it is tested in isolation and how it is tested.

Michael Lewis in his classic *The Lexical Approach* (LTP 1993) puts forward a strong case for focussing more on words than grammar in teaching, certainly in the early stages of learning. His approach is no longer new, but neither is it out of date. He does not go into testing as such, but parallels can be created.

## 1 Words and vocabulary

All the language we use is made up of words, but most BE trainers will see a difference, even if it is a subtle one, between *words* and *vocabulary*. This difference may not be so clear to all learners and is also often forgotten when setting tests. It is worth emphasising here, as the difference is of prime importance in a test claiming to test vocabulary.

A *word* in English is a group of letters separated by a space at the beginning and end between the groups of letters before and after it.

An item of *vocabulary* can often (although not always) carry meaning on its own and is not necessarily dependent on other words for its meaning. This is the difference between the two words *business* and *the*. An item of vocabulary can also be and often is more than one word, i. e. a lexical phrase. If learners know the meaning of *human resources*, do they need to know all the meanings and uses of the words *human* and *resources*?

Most BE trainers will recognise all this without having to enter into complicated explanations and a possibly rather academic discussion. However, in tests the difference between vocabulary and grammar often becomes blurred. This can affect the validity of a test claiming to test (BE) vocabulary. As with every test, the most important thing is to first clarify and define what you intend to test, in this case and the focus of this chapter, vocabulary in a BE context. Although there may be borderline cases, these should not be part of a test of BE vocabulary. Some examples of *words* which are not *vocabulary* are articles, prepositions and affixes. It is largely the vocabulary which carries the meaning or the sense.

**What is it?**

The terms *word*, *vocabulary* and *item of vocabulary* are used interchangeably in this chapter to mean the same thing.

### Validity and reliability

**Example**

In the sentence *I have work yesterday*, the word *yesterday* makes the meaning clear even if the grammar is wrong. Understanding the newspaper headline *Multinational fires top staff* depends on knowing the meanings of the words *fire* and *staff* in this context.

The questions which sometimes arise when differentiating between grammar and vocabulary may be one reason why language tests, and especially discrete item tests, often combine grammar and vocabulary in one test. These tests have titles such as *Use of English, Language Elements, Words and Structures* or even simply *Vocabulary and Grammar*. Even if the test, or section of the test, is not labelled at all or is simply referred to as a test of BE, in order for the test to be valid, the test setter must know what is being tested. Each item should be thought about in this way as well as what role vocabulary plays in the test and what weighting is given to the testing of vocabulary as a proportion of the whole test.

▶ Chapter 6

▶ Chapter 13

## 2 Vocabulary and grammar

Knowledge of vocabulary and grammar are what are known as enabling skills, that is they play a part in all aspects of language, including the four skills mentioned in the four preceding chapters. Knowledge of and ability to use grammar and vocabulary appropriately are not skills in themselves, but if learners have this knowledge and ability, they can produce language and use this to communicate effectively. Vocabulary and grammar are often the focus of BE tests. It is for this reason that they are addressed specifically.

One example of the distinction between vocabulary and grammar is can be seen in this test item:

**Example**

*Complete the following sentence:*
Jim is a hard worker. He works very _____.

If this is a completely open item, a test-taker may write any of the following: *well, effectively, carefully*, all of which are grammatically correct and show that he/she is aware of the correct grammatical form.

However the difference between these possible answers is one of vocabulary. *Hardly* is wrong, but is it the grammar or the vocabulary which is wrong? Changing the instruction to read *Transform the adjective into an adverb* makes the item into one clearly testing grammar and not vocabulary, even if what we really want to know is whether a learner has learnt what the word *hardly* means and can use it properly. (Also the word *transform* in the instructions is inappropriate and confusing as the word *hard* is not being transformed in any way.) The difference between *hard* and *hardly* is something which is difficult to test in a discrete item test, although a common mistake in spoken and written English. Knowing the difference is a matter of understanding the meaning and not the grammatical form, which makes it part of a vocabulary test.

The confusion between grammar and vocabulary in tests is one example of the misconceptions which occur in tests but rarely elsewhere. Other misconceptions in tests concern decisions on what language is correct in a given context, the importance of grammatical accuracy in spoken or written production, and the concept of native-speaker correctness.

## 3 Why test vocabulary?

The simple answer to this question is that we want to know whether learners know certain words and can use them accurately or appropriately. Although it is possible to test vocabulary explicitly in a speaking test, for instance by asking test-takers to produce words or describe pictures, this is usually inappropriate above a certain level and does not fulfil several of the criteria for tests of spoken production or interaction, especially if the focus is on the vocabulary rather than speaking. Practical considerations also mean that it is rare that vocabulary is tested explicitly in tests of speaking.

However, one way of testing particular vocabulary in BE contexts is to ask test-takers to describe graphs or diagrams orally or in writing. To do this they are obliged to use particular vocabulary items, such as *raise* and *rise*, *drop*, *fall*, *trend* and *tendency* accurately.

In general, tests of vocabulary involve reading and writing. Test-takers may be asked to recognise, identify or choose a word from several possibilities or to produce a word in writing after being given a prompt. In the first cases, receptive knowledge is being tested, and in the second (receptive) knowledge is being activated. In both cases we can assume, if the test fulfils the criteria for a good test, that the test-taker *knows* the word at some (however low) level.

This does not preclude guessing, whether intelligent or not, which firstly can never be totally eliminated from tests and secondly is (or should be) a part of strategic language use.

## 4 Aspects of vocabulary

The various ways in which vocabulary can be defined are a continuum, starting with the "bare" meaning of a word or phrase and going on to the use, the different forms of words and their collocations.

### Meaning

The meaning of a word can be tested by asking for a translation or a definition or by providing a translation or definition and eliciting the word or lexical phrase. Although it is doubtful whether this gives you any indication of the test-takers' level or skill, it can be acceptable as long as you are sure about which items you will accept as correct and can explain this to the test-takers. You will also have to decide how far the word has to be absolutely correct (e. g. the spelling) to be awarded full points. A definition like "a place where you keep money" where the test setter's intention is to elicit the word *wallet* or *purse* could just as easily produce the word *bank* or *moneybag* or even *sack* or *under the mattress*. The definition can be narrowed down by making it more specific, by providing a sentence with a gap or by restricting the answer to one word. However the more context is given, the greater the reading load in the test will be. As well as this, at more advanced levels there is relatively little vocabulary for which the meaning can be defined precisely, and so it may only be possible to test a few items of vocabulary in this way. Meanings of words can also be tested using multiple choice questions where a word in a context is defined in three (or four) different ways and the test-taker asked to choose the correct one. All the guidelines for setting multiple-choice questions apply in this case.

▶ Chapter 13

### Use

Whether test-takers know how to use words appropriately can be tested by, for instance, asking them to write a sentence using the word. This, while a fairly valid test of use of vocabulary, has two disadvantages. One is that the sentences produced may be very simple, and although not wrong, not show clearly that the test-taker knows the meaning or the specific way in which the word is used. The other is that deciding on the criteria and marking tests of this type will be time-consuming. For instance, you will have to decide how far parts of the sentence which are not the word being tested are taken into account, i. e. how far do these have to be grammatically and orthographically correct? In spite of these considerations, questions of this type will provide an interesting and fairly valid picture of the test-takers' ability to use vocabulary. This type of test is most useful in diagnostic tests and where the test-takers know what is expected of them.

**Remember**

It is always best to let test-takers know what is expected of them and how they can get the best results in the test.

### Transformation

Word transformation is often tested at advanced levels and is also a valid way of testing whether learners know how to use items of vocabulary in various contexts. If the test-taker transforms a word correctly, it can be assumed that he/she knows the word, but of course intelligent and experienced (and even desperate) test-takers will use guessing techniques and some of the time may get it right. Word transformation tests are fairly straightforward to mark as the range of answers given will be fairly small.

**Example**

*Fill the gap with an appropriate form of the word in brackets.*
It was impossible to read his handwriting. His handwriting was _____ (legible).

In general in these tests, only absolutely correct answers are acceptable.

## Vocabulary in context

The question of context is clearly important as soon as we go beyond the simple meaning or form of the word. The more context is given, the easier it will be for the test-taker to produce the word the test setter is looking for. However a great deal of context for one item of vocabulary increases the reading load and the time needed for the test without necessarily producing a significant improvement in the usefulness of the result. Vocabulary is often tested by using texts with gaps, but although this gives the test-takers a context for the vocabulary, vocabulary tests using long and complex texts with gaps are extremely time-consuming to mark. A single context for all the items to be tested may not be available, and using too many different texts and contexts in a test requires flexibility on the part of the test-takers, as they have to adjust to a new context with each new text. Two different contexts for texts are enough in most cases. Remember that the test-takers have to be able to understand most of the other vocabulary in the text in order to be able to fill in the gaps, so that some of the issues raised about tests of reading apply here as well.

▶ Chapter 7

If the answers are straightforward or obvious, then the test is often too easy. A test-taker may change the context by putting in a different word from that intended by the test setter and thus change all the other words which are filled in later gaps in the text. The question of spelling and absolute correctness also applies to gap-filling tests. In general inexperienced test setters or those with limited time are advised not to use gap-fill tests for vocabulary as these often cause more problems than they solve and do not provide useful results.

For these reasons, multiple-choice items are often used in tests of vocabulary. Items of this type can test knowledge of the meaning and the use of words or lexical phrases. Although the test-taker may have to adjust to a new context with each item, this is short and should be relatively simple. However it is not always easy to think of enough distractors for the items. In general, for tests of vocabulary, this makes 3-part multiple-choice items better than 4-part. The distractors must be near enough to the correct answer to be plausible but not so near that they could be correct. The distractors may be chosen because they are near the correct answer in meaning, but do not collocate well in the context as in the following examples.

**Example**

All the managerial _____ in that department are highly qualified.
a) humans
b) personal
c) staff

If you agree to our prices, we can _____ the deal and sign the contract.
a) close
b) end
c) finish

It is often better to tell the test-takers to choose the *best* word rather than the *correct* word in this case.

Multiple-choice questions can also be used with easily confused words, although it is not always easy to find more than two options, which makes the test generally too easy. Sometimes distractors are chosen because they look like the correct word, although this may actually have the effect of confusing the test-taker and can make the test unfair.

Test setters should also consider how precise they want to be. In vocabulary exercises which are done in class, a trainer can discuss why only some answers are correct or better than others. In a test this is not possible. In view of the spread of English in international contexts as well as the general issue of communicative effect, it is often not easy, although by no means impossible, to define what the right answers are in a vocabulary test. In particular in international business contexts there may be fuzzy notions of what is right and wrong and what is acceptable. We should not let our personal notions of correctness affect the way a test is set.

There is also vocabulary which is easier for speakers of some languages than of others. False friends are different from language to language. If all the test-takers are from one language group and all make similar vocabulary mistakes, this can affect the choice of items to be tested as well as the choice of distractors in multiple-choice questions.

## 5 What does it mean to know a word?

A test of vocabulary tests whether test-takers know specific items of vocabulary. Having established the difference between words and vocabulary and then the difference between vocabulary and grammar, we must decide what specific knowledge of vocabulary and then what specific vocabulary is to be tested. Is it enough if a test-taker can recognise a word? Should they recognise the word in isolation?

In fact, apart from very specific and mostly concrete words for objects, it is usually hard to know what a word means without any context whatsoever. This may work at low levels, but as soon as a learner knows several contexts or several meanings for a word, it will be difficult to test words or lexical phrases in isolation.

Does it matter why the test-taker recognises the word? Is the test intended to see if test-takers can use words in an appropriate context? Is the context spoken or written? Should they know several ways of using the word or just one? Should they know a range of words and how wide should this range be?

It should be borne in mind that the range of vocabulary actually used in English is very small. Estimates say that there are roughly a million words in English. This is difficult to judge, however, as many of these words do not carry meaning, i. e. are not vocabulary in the sense meant above. An educated native speaker probably knows what about 20,000 of these mean. The average native speaker or competent user of the language uses about 2,000 words regularly. Of these about 1,000 are used actively about 80 % of the time.

This is also the case for BE, although specific vocabulary for particular fields may be used more often. It is usually only necessary to recognise obscure vocabulary. Active use of it may not be important. In fact, this can often sound pompous or affected.

Knowing a word is more than just knowing its immediate meaning and being able to provide an acceptable translation of it. It also includes being able to change it according to the sentence structure, to use it in several contexts and/or to use it to make sense in English. It is important to clarify this before you start setting a test. Whether or not you label the test as one of vocabulary, to be valid it must be testing what you intend.

It is of course quite possible to guess or deduce the meaning of a word correctly from the context. This is one of the strategies which successful users of language employ and an important one in communication and effective language use. However it is extremely difficult to ascertain the effective use of strategies in a language test fairly and reliably.

The tester can never be certain if the test-taker already knows the word or if he/she does not know it and is using a strategy to work out its meaning. Deductive strategies can only be used and tested with words in context. It is not possible to guess the meaning of a word without a context.

The most valid tests of vocabulary are those testing receptive knowledge of the vocabulary item and whatever aspects of it the test setter chooses. Whether this knowledge is also productive can only validly and reliably be tested in a test of language production, spoken or written. This means that setting a test of vocabulary can be fairly easy. However, the results of the test cannot be used to make any inferences or statements about the test-takers' productive use of language. Knowing a wide range of vocabulary will not necessarily mean that a test-taker will be able to interact effectively or to write coherently. Nor does it mean that the necessary vocabulary will be used spontaneously and appropriately in a variety of different contexts. All these factors mean that it can make sense to test vocabulary separately. In most cases a test of vocabulary will either be a small low-stakes test (in the classroom) or part of a larger test of skills and differentiate between receptive and active use of vocabulary.

## 6 What vocabulary should be tested?

**Remember**

Although meaning can be tested by asking for a translation, it is doubtful whether being able to translate a word is any indication of knowing how the word is used.

The specific vocabulary to be tested will depend on the context. Tests of vocabulary in BE contexts are carried out to see whether the test-takers can use specific vocabulary from their field accurately, and (often) whether they know the meaning of "false friends" and can use these properly without causing confusion. It is usually use rather than meaning that is being tested.

Only those vocabulary items can be tested which test-takers have learnt or should have learnt. In a placement test these will be words they are expected or need to know. In an achievement test, the vocabulary will be taken from the course content, so the learners should know what vocabulary they will be tested on.

If you are setting a test of vocabulary, it will be necessary to define exactly which words you want to test, where they come from and what they mean, bearing in mind the above. In general English, at elementary levels, it is fairly easy to define which items of vocabulary and which meanings test-takers should be expected to know. With BE, the situation is slightly more complicated and vocabulary less easy to predict. Learners who are productively at a relatively low level may have an in-depth passive knowledge of the vocabulary of their domain of expertise. Whether in general English or BE, as learners become more advanced, the range becomes too great to be

able to define easily. This is somewhat compensated for if a specific domain is being tested, i. e. business, and becomes easier the more specific the area is, e. g. marketing. It is probably best to concentrate on the specific words that you expect or want test-takers to have learnt and assume that the other vocabulary used in the test will be familiar. Even if certain words are not the direct focus of the test, they will play a part in defining the context and so will be tested indirectly. Thus what is said about ease of reading texts applies equally to the contexts given in vocabulary tests.

▶ Chapter 8

## 7 Factors affecting the testing of vocabulary

The following factors all only affect the testing of vocabulary separately. Testing knowledge of vocabulary in tests of skills such as writing or speaking should be approached differently.

▶ Chapter 9

In a vocabulary test, it must be determined how much other knowledge, e. g. of grammar or other vocabulary, plays a part in finding or choosing the right word in a context. The test setter should also decide whether receptive or productive vocabulary is being tested and how far the vocabulary has to be absolutely correct in every respect. Knowing a word can mean different things. It may mean recognising it, but not knowing its meaning or it may mean knowing what a word means in isolation but not knowing how to use it in a context.

▶ Chapter 10

### Validity and reliability

For a test of vocabulary to be valid and reliable it must contain enough items as well as enough from the basic vocabulary the learner is expected to know. Very obscure words and phrases should only be tested in isolation for good reasons. These may be that they belong to the test-taker's particular field of business or it may be to differentiate between strong and weak test-takers, e. g. in a placement test. However test setters should consider whether the knowledge of obscure words is important enough to be worth testing.

**Remember**

Intelligent guessing plays an important part in language use and helps both with communication and the language learning process.

As in real-life use of language, guessing will play an important part in tests of vocabulary, especially in closed-question types. Whether this is to be penalised in some way will depend on what the test is for.

In order to be able to activate a word, a stimulus is necessary and it is this that the test provides. The test-taker decides on the form of the stimulus – sentence or text with a gap or gaps, translation, definition, etc. However what the test setter cannot be sure of is that the stimulus will only produce the word intended. This can mean that several answers are possible. If a particular word is expected, it may be necessary to change the stimulus and even the type of test. It definitely means that the test should be tried out on several people to ensure that the answers given are the ones intended. It may also mean that there is no alternative to accepting a range of answers as correct. The question of spelling (and if vocabulary is tested as a part of a test of speaking, pronunciation) also has to be considered. In many cases spelling is accepted as correct if the word is recognisable and the incorrect spelling does not cause confusion, e. g. as with *personal* and *personnel*.

## Test formats

Discrete item testing of vocabulary is best in placement tests and in those cases where the range of vocabulary is clearly defined, e.g. in classroom achievement or progress tests. However, there will probably be some vocabulary which cannot be tested effectively in discrete-item tests and may have to be taken into account in tests of skills. There are two main reasons for this.

1 The real choice the test-taker makes is usually only between two possibilities (e.g. *make* and *do*). This means that a three- or four-part multiple-choice item will not work as the real choice is only between two of the options. The others (e.g. *take* and *put*) will either confuse (weaker test-takers) or be ignored (by stronger test-takers). It is often not easy to find more than two distractors for multiple-choice vocabulary items, which can lead to multiple-choice tests of vocabulary including words which look similar but have no similarities in meaning.

**Example**

His knowledge of the pharmaceutical marketing is very _____ ; he knows a lot.

a) brave
b) breadth
c) brief
d) broad

The options have only been chosen because they begin with *br-*, one is not even an adjective.

2 There are many words which would be appropriate. This is particularly the case with gap-fill items. If the example above were a gap-fill item rather than a multiple-choice question, possible answers could be *extensive*, *great*, *far-reaching*, etc. There is no way of ensuring that the test-takers write what the test setter intends. This is sometimes dealt with by providing the first letter of the word intended, but this can unfairly penalise a test-taker who knows an equally valid word starting with a different letter. Using multi-matching gets round this to a certain extent but poses similar problems to those in multiple-choice items. Multi-matching tests are those where gaps in a text can be filled by choosing from a set of items in a list. The list contains either the same number of items as there are gaps in the text or more than this number. The words must be similar enough to be plausible but different enough so as not to fit in more than one gap.

## Accuracy and range in speaking and writing

In higher-stakes tests, the best solution is to use different ways of testing vocabulary and include vocabulary as one of the criteria in tests of skills. Although a small number of discrete items may be possible for vocabulary testing, it is probably more effective to test vocabulary use and range as part of a test of reading, writing, listening or speaking as it forms a natural part of these tests. There may be a case for not testing grammar specifically in speaking or writing, but as vocabulary carries meaning, it is essential for effective communication. In tests of BE writing, the correct or appropriate vocabulary will also involve issues of style and register. This makes it easier to determine which words or phrases are best or correct in a given context and makes a test of vocabulary easier to mark. Active knowledge of vocabulary can be tested in tests of writing and passive or receptive knowledge in tests of listening and reading.

■ Testing vocabulary is fairly straightforward and should not be made unnecessarily complicated. Bearing in mind the limitations described above and the importance of vocabulary in communication, it is best not to concentrate too much on one type of test or questions but to consider the role vocabulary plays in all skills and areas of language use. Although this chapter has concentrated on the testing of vocabulary separately, much of what is said is relevant to tests of skills of reading, listening, speaking and writing. Defining the vocabulary the test-takers should know will help to make the decision on what and how to test. ■

### To think about:

- What vocabulary do you expect test-takers to know?
- Do you want to test vocabulary separately or as part of another test?
- Are you testing productive vocabulary use or receptive vocabulary knowledge, range of active or receptive vocabulary or accuracy of use?

CD-ROM

*"Finally, the assessment question – how do we know that grammar learning has occurred? What claims are we asserting about the learner's grammatical knowledge on the basis of assessment?"*

Purpurra, James (2004): Assessing Grammar. CUP, p. 24

# 12 Grammar

## Grammar as the skeleton of language

- What are the different views of grammar?
- Should grammar be tested in isolation?
- How can grammar be tested?

If grammar is the skeleton of language, this can mean that it is essential for language to remain in existence at all, or it can mean that it is useless without the flesh and bones. Both these views reflect the fact that the role of grammar in teaching, and by implication in testing, has been controversial for some time. It seems that grammar cannot be ignored in any discussion of language training, but what part it plays exactly remains to be defined. BE trainers, like all other language teachers, will have varying views on the importance of grammar. These will be reflected in their teaching and also in the way they test grammar or if it is tested at all. There is more controversy about the role of grammar than about any other of the aspects dealt with in this book, and yet grammar may be the easiest to test in isolation. However this should not lead to it being the focus of BE tests. Knowledge of and ability to use grammar enables language users to communicate their message, but grammar should not be allowed to overshadow all other aspects of language use. A test of grammar is not a language test, even if many language learners may see it this way. Of the over 50 descriptive scales in the CEFR only one is concerned explicitly with grammatical correctness. This should be remembered when setting a language test, which means that there is a better case for leaving out the testing of grammar completely than for making grammar the focus of a test for a particular CEFR level. The decision whether, and how much, to test grammar will depend on the situation, the learners and

the test-setter. This chapter will look at how grammar can be tested, whether in isolation or in connection with a language skill.

## 1 What is grammar?

Grammar can be defined as the structure of a language, without which it is not possible to link words, ideas and concepts. There is no doubt that grammar continues to play an important role in teaching and also in BE teaching, although it is probably the case that its importance has diminished somewhat over the last few years. In many cases books claiming to teach or practise business grammar are usually only doing the same things as other books of grammar exercises but just using BE situations and vocabulary. There is probably no such thing as *business grammar* i. e. grammar used exclusively within a business context, but it may be the case that certain forms and structures are used more often in business contexts than in others. Most BE courses contain some grammar exercises and, as BE is usually taught at advanced levels, may contain more advanced grammar. Advanced grammar often means grammatical structures which are little used and without which learners can manage perfectly well. These include things like the causative, the past perfect continuous and inversion. It is doubtful whether there is any need to teach or test the active use of these, although recognising their meaning may be important in some contexts.

## 2 Views of grammar

Tests of grammar are often given names such as *Use of English, Structure and Expression, English Usage*, among others. Some of these names are also used for tests of vocabulary. They usually test knowledge of the system of English, and sometimes, but not always specifically, a mixture of grammar and vocabulary. The name of the test may say something about the view of grammar underlying the test. Whatever it is, your own view of grammar will influence the content and the method of your BE teaching. In turn this will probably influence your view of how grammar should be tested in a BE context. You may feel that grammar is only useful in as far as it contributes to meaning and the communication of this meaning and has no intrinsic value otherwise. Your learners may only be interested in using English in business contexts, and not in the rules or the reasons for the correctness or otherwise of their language. If this is the case, you may not want to test grammar as such and may wish to test whether and how it is actively used by test-takers, i. e. in tests of speaking and writing. If you never treat grammar as a separate subject in class, but only as an enabling element of language, there will be little point in testing it in isolation. If, however, you spend a lot of time in class doing grammar exercises and your learners are used to this, then it may make sense to test grammar knowledge or ability explicitly. Just as you decided what your view of BE is at the beginning of this book, you should decide how you see grammar before deciding where and whether to test it in a test of BE.

Your view of the importance and role of grammar will probably also depend on whether you are a native speaker of English or whether you have learnt English as a foreign language and how you did this. If you are a native speaker, you may only have learnt about grammar in order to become a BE trainer and feel it is less important in communication. If you have learnt English as a foreign language in a country where it is spoken, your views may be similar, whereas those who have learnt English at school and

university may judge the importance of grammar differently and possibly consider grammatical correctness more important in language use.

## Grammatical knowledge

Some approaches consider that learners should learn and internalise grammatical rules. It is certainly useful for many learners, especially those who have been through a school system where language was taught by means of grammatical rules. However, as all language teachers know, knowing the rules of grammar is no guarantee that they will be applied correctly when speaking or writing. Rules may be useful for some learners; others may not find them interesting or helpful and may prefer to learn only from examples and practice. Many learners feel more comfortable when they are given rules, even if they cannot always follow them in practice. If you or your learners feel that grammatical rules are useful, then you will probably give these and also do exercises in class which practise the application of the rules. This will give you a justification for testing whether learners have internalised and can apply the rules of grammar.

Rules may include at a basic level, e. g. the third person singular 's' or the formation of the passive, the use of prepositions for times, days, months, years, or the use of the present perfect as opposed to the simple past. Knowledge of these rules does not however imply that the test-taker will be able to use the language appropriately in a given context. Testing knowledge of grammatical rules would mean asking test-takers to give the rule for using e. g. a certain tense or preposition.

In general, language tests (and especially BE tests) no longer test grammatical knowledge, i. e. explicit knowledge of grammatical rules. There are many good reasons for this, but only one will be mentioned here.

Exceptions to rules are often forgotten, even when there are as many of them as instances of the rule itself. It has often been forgotten that teaching and testing rules also includes teaching and testing exceptions. Many things which are considered exceptions in the strict grammatical sense are in fact so commonplace that they are heard more often than the use of the rules. One example is the rule that stative verbs cannot be used in the continuous form. The number of exceptions to this rule means that many learners find it easier to learn examples rather than the rule and the exceptions.

**Remember**

There are BE course books which use grammatical terms explicitly. If you do this, learners will be familiar with them. This will influence the choice of words in your instructions for a grammar test as you will be able to use grammatical terms in these.

▶ Chapter 14

Even though it is unlikely that knowledge of grammar will be the focus of a BE test, when it is, exceptions are sometimes forgotten or ignored. This should not happen, but neither should exceptions be allowed to become the focus of a test of grammar. Everyone knows how long it takes some learners to use the third person singular 's', even when other aspects of their language, such as vocabulary or even other grammar, can be considered advanced. This is not because they do not *know* it or *cannot* use it and probably most test-takers at any level above beginners would get full points in a test of only this. There are many explanations for learners' seeming inability to use the third person singular 's', including that it carries no meaning, so is not important. Whatever the reasons for this phenomenon, testing knowledge of it would provide no usable results. The same may apply to testing other aspects of grammatical knowledge.

## Grammatical ability

This means the ability to use the knowledge referred to above. A test-taker may know a rule without having been taught it or even being consciously aware of it. Knowledge of grammar needs by no means be explicit. It could well be that a test-taker knows a rule for using a certain preposition, for instance, but does not express this knowledge in terms

of the rule, but by means of an example or simply by using the preposition correctly. This approach is far closer to the approach generally used in BE tests of grammar. Most of what we usually think of as grammar tests are tests of grammatical ability. They ask test-takers to apply the rules, without asking them to give the rules. This ability is however the most difficult to define. A grammatically correct sentence may convey a completely different meaning from that intended.

**Example**

This can easily be seen in a confusion of tenses, for instance, or of words like *boring* and *bored*. What do the following sentences mean and is this what the speaker intended?
All the participants in the discussion were bored.
All the participants in the discussion were boring.
I have considered your application.
I have been considering your application.
I am considering your application.

Teaching the form of the past perfect continuous, or any tense, does not mean that learners will know how or when to use it or even if they will use it at all. Especially in tests of writing, parts of what is written may be grammatically correct, i. e. the form is correct but the sentence is impossible to understand. It is important to identify what you mean by grammatical ability. Is it ability to produce the correct form, to use it in a context and/or to use the correct form to express a particular meaning? All these things can be the focus of a test of grammar.

## Grammatical performance

This is what is tested in tests of productive skills, i. e. speaking and writing. It is not the same as grammatical ability, which is more likely to be receptive or restricted to particular situations. In unfamiliar contexts or when skills are involved which are not as highly developed as the test-taker's grammatical knowledge or ability, he/she may make grammatical mistakes which may affect the message to be communicated. This generally means that the test-taker is not able to deal with the particular situation. This could for instance apply to particular forms of writing, such as academic writing or situations in a test of speaking, for example giving a presentation or chairing a meeting. It is also possible that a test-taker in a test of performance will make no grammatical mistakes as such, but will not be able to achieve a high level of performance because he/she does not have grammatical structures available to convey the necessary meaning. This may be the case with cautious language users who restrict what they say to what they can say and try to ensure that they only use correct grammar. It is important to think about not only whether what is said or written is grammatically correct but also whether it is appropriate and whether a suitable range of grammar is used.

▶ Chapter 10

## Metalinguistic knowledge

Metalinguistic knowledge means knowing the names for grammatical terms and has little to do with language ability or performance. It is connected with grammatical knowledge but refers to knowledge of grammar as a system of rules rather than particular knowledge of English grammar in use. Metalinguistic knowledge of grammar is unlikely to play any part in a test of BE.

## 3 Why test grammar at all?

There is a fairly strong case for not testing grammar explicitly when testing performance, especially speaking or writing. However this may make the testing of grammar in a separate part of the test desirable or necessary. If grammar is tested in a part of the test dedicated to it, this will mean that it can be ignored elsewhere and more focus can be placed on possibly more important aspects of language, such as interaction (in a test of speaking) or structure (in a test of writing). This is the strongest case for testing grammatical ability separately, i. e. to reduce its importance in tests of speaking and possibly writing. If it is included in tests of speaking and writing as one of the criteria (usually called *grammar* or *grammatical correctness or grammatical accuracy*) then its weighting must be considered carefully.

The question of weighting will of course also apply to any part of a general BE test which specifically tests grammar. Is grammatical ability (the aspect of grammar most usually tested in written tests) to be given equal weighting to spoken performance? The scores for different parts of a grammar test may also have to be considered. You may want to give fewer points for what you consider to be basic grammatical ability and more points for the ability to identify and use complex grammatical forms.

▶ Chapter 13

There are grammatical mistakes which make little or no difference to the communication of the message, in speaking and, in many cases, even in writing. Many of these are the mistakes often seen in public notices, signs etc. and which are made by non-native speakers. Two examples are *I suggest/recommend to go* instead of *going* and the wrong use of *by* and *until*. *Send in your application until January 15th* is a sentence commonly found in job advertisements. In both these cases, and many more, the mistake makes no difference to the communicative message. Many, if not most, non-native speakers would not even notice it and many, if not most, native speakers of English would not recognise it as being important – unless they are teachers of English! However, even if there is no difference in the communicative effect of what is said or written, you (as a BE trainer) may feel that it is important that learners know and can use (at least in theory) particular structures and forms correctly. This is also a good argument for testing grammar.

**Remember**

To be valid, a test has to test what it sets out to test. Your definition of grammar will be the basis for the validity of your test.

In the case of the mistakes mentioned above, the logical consequence would be not to penalise test-takers for making them in a performance test. On the other hand, if results in a test of grammar show that test-takers have not learnt or internalised a particular structure or form, they can be useful for assessing learning progress, for placement and for course design. This is one argument for testing grammar in several types of test, such as diagnostic, placement or achievement tests.

In BE, correct grammar may play a more important part in the impact of and impression given by business communication. This is however often more a question of style rather than grammatical accuracy (which may be picked up by a spell-checker anyway) and so better tested in tests of writing rather than discrete item tests of grammar. If you wish to test it separately, sentence transformation can be one effective way of doing this.

Grammar, or grammatical ability, can be tested on paper to set a basic level for a course so that you know where to continue. This would be the justification for including grammar in a placement or diagnostic test. Although the results of a grammar test, like those of a vocabulary test, give no indication of the test-takers' performance ability, grammar tests are fairly easy to set, quick to mark and often expected by learners and institutions or companies.

## 4 Grammar tests and exercises: the difference

Tests are not set in order to allow learners to get practice. This is the basic difference between a test and an exercise. A test should reflect an individual's performance and ability at a particular point in time. The results must be useful for placement, diagnosis, to see whether the learner has achieved a set of goals or for a forecast of proficiency. Nonetheless, many so-called tests of grammar set by teachers are in fact simply exercises. They consist of a list of questions, all testing one structure, e. g. with instructions such as *Put the following sentences into the passive*. It should be clear by now that these do not fulfil the criteria for good tests and so are not useful for various reasons. Tests (in this case, of grammar) should reflect a range of ability or performance and should allow the test-takers the possibility of achieving the best results possible. They should be authentic and have criteria against which they are marked and scored.

Grammar exercises, of the sort found in course books and done in class, are intended to help learners to acquire, but more importantly, to practise certain structures. While they should also be authentic, they are designed to give learners examples of different ways in which the particular structure is used. If you want to test whether test-takers can use the passive appropriately, one question on it is probably enough. One grammatical structure cannot form the focus of an authentic test of grammar. The results will only be useful if test-takers answer all the questions correctly. If this is not the case, the results will probably be meaningless. If a test-taker makes a mistake in one question, but gets another right, what does it mean? Does he/she know the structure? Can he/she use it properly or not? A test of grammar must test a range of different structures appropriate to the level of the test-takers and to the sort of BE they have learnt, in order for the test to be valid, reliable and authentic. This means that exercises from books cannot be used in tests. It is also the main reason for using question types like multiple-choice questions, as they can be used to test many different grammatical forms and structures. Exercise forms are usually specifically devised to practise one particular structure and may not be suitable for practising or testing others.

## 5 What grammar can be tested?

The choice of what grammar to test will depend largely on your situation and the type of test you are setting. For a placement test, you will have to decide what learners in each group should know already and what differences there are, in terms of grammar, between the different courses. In an achievement test, you will want to see if learners have learnt the grammar which was taught in the course (this is still no justification for using exercises!) and in a proficiency test, it will be important to identify what grammar is necessary for the intended performance. This decision, on what grammar to test, will be the basis for the validity of the test, whether this grammar is essential or useful for the learners in their business context.

In order for the test to be valid and reliable, the choice must be a realistic one. Even at an advanced level, there are grammatical forms which are not used often, so if you wish to test these, the test-takers' chances of passing or failing should not depend on them. This is one important reason for not only using a test of grammar to determine a level or as the basis for entry to a course.

▶ Chapter 2

If your test is a test of performance in BE or a proficiency test, and if it is to be linked to a CEFR level, it may not be necessary to test grammar separately at all. You may feel that grammatical ability is important and should be tested separately, but find that it is no indicator of performance, or even that it is quite possible to perform very successfully and even grammatically accurately with a fairly narrow range of grammatical structures.

Although, in general, BE teachers feel that being able to use conditionals correctly and accurately is essential for conducting negotiations, it has been suggested that this is in fact not the case and that basic grammar is quite enough.

One reason for testing grammar explicitly and separately is that you are requiring learners to demonstrate their knowledge and ability, which can only be done with difficulty in a performance test. It is neither possible in a test to force test-takers to use language accurately nor is it possible to force them to use certain structures. Attempts to do so will only make a test inauthentic and lacking in interactivity. Even a question like *What were you doing when (the storm started)?* which most teachers would agree requires an answer using the past continuous can be answered with any of the following *I can't remember exactly/I'm not sure but I think I had just eaten breakfast/I think I just wanted to make a phone call*. One way around this (apart from instructing the test-taker to use the past continuous in the answer!) is to provide options to choose from, which, again, does not test active production of the structure.

▶ Chapter 3

Whatever grammar you choose to test, it is important to include a range and also to include some supposedly easy or basic grammar, so that you will get a reliable picture of the test-takers' grammatical ability and not overemphasise this either in the test or in the results.

## 6 Factors to consider when setting grammar tests

As mentioned above, tests of grammar are fairly easy to set. The most common type of question is multiple-choice, although open-ended questions and sentence transformation are also used. The reason for using multiple-choice is precisely because it is a question type which can be used for a wide range of grammatical structures. It is also relatively easy to think of three or four plausible distractors for multiple-choice questions testing

grammar. Unlike multiple-choice tests of vocabulary, however, where the distractors are often completely implausible, in grammar questions the difficulty is often that several distractors are possible. This means that details of the context have to be provided for each question. Tests of grammar sometimes use one text with gaps and multiple-choice possibilities for the gaps so that test-takers do not have to internalise too many different contexts in one test. The drawbacks are those which apply to all multiple choice questions.

Proof-reading, if the test focuses on grammar mistakes is also an effective way of testing grammatical ability, but can only be reliable above a level where test-takers can identify mistakes.

▶ Chapter 3

Whatever form you use, make sure that the test-takers know exactly what to do. This is particularly important in open-ended questions or sentence transformation questions where you want them to use a minimum number of words in their answers. If the task itself is complex, the interactivity of the test will be low. It is often helpful for test-takers to be given an example (which can be at a much lower level than the test itself) and also to have practised the type of question beforehand.

It is also important that the structures and contexts you use are authentic. If you cannot find an authentic example of use for the grammar you want to test, then it is probably not worth testing. The context should be clear and the language natural, whether spoken or written. Make sure you are not testing something just for the sake of testing it or insisting on language which is rarely encountered in real life. In fact 20% of grammatical structures and forms are used 80 % of the time, so it is probably more worthwhile focussing both teaching and testing energy and resources on these.

**Example**

An example of this would the use of *whom*, which, while entirely appropriate in the context of *To Whom It May Concern* is rarely used elsewhere nowadays.

There are also examples of so-called grammatical rules which appear to be constantly broken by native speakers. Many teachers teach rules for using the various ways of expressing the future in English and yet, although apparently the rule for expressing uncertainty about the future prescribes the use of *will*, all the options given in the following question could be used quite correctly (not only by native speakers).

**Example**

I am not sure what he _____ next Wednesday.
a) is doing
b) is going to do
c) will do

Another example for a different rule is the following:

**Example**

My daughter is in Hawaii at the moment and she _____ there.
a) does love it
b) is loving it
c) loves it

The more learners are exposed to authentic English, the more they will hear examples like this. This makes it unfair to expect something only theoretically correct from them in a test. Both when choosing distractors for multiple-choice questions or deciding

which answers to accept in open-ended or gap-fill questions, the varieties of English and BE which exist must be taken into account. This not only involves British and US English but also other varieties which learners will encounter. Particularly as far as grammar is concerned, there are many versions which are acceptable, and test-takers should not be penalised for using or for choosing these.

As correct language is not only a question of form but also one of meaning, test setters should be careful not to concentrate too much on form to the exclusion of meaning. In many exercises, forms are tested although they are not often or ever used like this in real communication. One example is the use of the imperative, rare in spoken English except in very specific situations. Using the imperative *Open the window!* can convey an inappropriate or wrong message under normal circumstances and is by no means a polite request. It is better here to use a polite form wrongly. *Could you please opening the window?* is therefore, while grammatically incorrect, better.

The pragmatic meaning of language, while becoming more and more of a consideration in teaching material, is still not always taken into account in tests of grammar, which tend to focus on form. There is nothing wrong with this focus, as long as the language used and tested remains authentic. One way of doing this in tests of performance where grammatical accuracy is one of the criteria is to only consider this as the final criterion when marking. The criterion description can be given last (after, for instance communicative impact, structure and vocabulary) and can also be made explicit in the instructions to the raters. Raters can be told only to mark grammar after they have awarded marks for the other criteria. It can also be made part of rater raining sessions.

■ Whether grammar is tested in isolation or as a part of a test of a skill or in combination with for instance reading, listening or vocabulary, it is important to remember that grammar knowledge and ability is rarely, if ever, the main point of language use and may play a relatively minor part in authentic communication. So, although tests of grammar are relatively easy to set and can fulfil the criteria of validity, reliability and objectivity, in order to fulfil the quality criteria for good tests, they must also be both authentic and interactive. ■

CD-ROM

### To think about:

- What is your view of grammar?
- Why do you want to test grammar?
- Do you want to test it separately or as part of another test?

*"The decisions we make on the basis of test scores can affect both individuals and programs."*

Bachman, Lyle F. (2004): Statistical Analyses for Language Assessment. CUP, p. 39

# 13 Marking, scoring, criteria

## On mistakes and correctness of language and its evaluation

- What are mistakes?
- How is performance marked?
- How are results calculated?

Testing is more than just setting questions which are then answered by test-takers. Even professional test developers, who set tests full-time, do more than just think of the test questions. Test setting also includes deciding on the answers expected, on the number of points to be awarded for each answer, what the total score will mean and if the final result is a pass or a fail. This involves not only deciding on the total number of points to be awarded for a test, but also on how points will be deducted from this total, i. e. which answers are right and which mistakes are important. It also means deciding who will do the actual marking in the test and what the marks will mean for the test-taker.

All these decisions are particularly important if a test or a test format is to be used several times. It will not be possible to compare tests which are supposed to be testing the same thing if they are marked, scored and graded in different ways each time. If the same test is marked and graded in different ways or using different criteria, it will be neither reliable nor valid.

In some cases, the actual number of points may not be particularly important, as the total will be expressed as a percentage. A percentage is probably easiest for test-takers to understand.

## 1 Mistakes

Teachers (and BE trainers) sometimes seem to see their main function as correcting learners. They spot mistakes, point them out, correct them and try to teach learners not to make them again. It is often implied that there is some accepted definition of what a mistake is. In teaching, the decision on what a mistake is can be left to the individual trainer, but in testing it is essential that those involved are working with the same definition in order for the scores to make any sense. Before deciding on criteria and scores for tests it may be helpful to think about and discuss the following questions.

Trainers' expectations of the language which can be expected and how free of mistakes this should be will vary enormously. It will be important to decide on this, even if one person is setting and marking the test. Whatever level the test is at, the highest score should be defined. Unrealistic perfection should not be expected. It may be useful to consider how you yourself would perform in the test. Even trainers would probably make some mistakes in a test of production, although they would be of quite a different nature from those made by learners. An examiner's/rater's role is to assess, but this means more than just finding mistakes. The attitude trainers, examiners and raters have to their role in the test will affect the scores they give.

**Remember**

Errare humanum est.

**Did you know?**

Mistakes which are unacceptable to teachers or to some native speakers may be perfectly acceptable and not even considered as mistakes in other (native speaker and non-native speaker) contexts.

The type of mistakes made in different tests and their acceptability will also vary. To some extent, this will depend on the level of the test-takers and the test tasks. Before beginning to think about criteria, define the type of mistakes which can be expected, those which can be accepted and those which are unacceptable. The mistakes which an individual trainer finds unacceptable may correspond to those he/she would definitely correct when teaching. Acceptable mistakes may be those made when the particular language aspect has not been completely assimilated by the learner, not practised enough or not learnt at all. This again may depend on where and when the mistake occurs as well as on the skill being tested, the type of test and, again, the level of the test-takers.

**Remember**

Few sentences are complete in native-speaker spoken language!

Mistakes may be those of commission or of omission. Using a word wrongly may be considered a more serious mistake than avoiding a subject where the vocabulary is unfamiliar. Not answering a question at all may be worse than giving a wrong answer using correct grammar. Answering in one word may be acceptable or it may be rude. Mistakes may be made in grammar, vocabulary, register, pronunciation, intercultural communication as well as in other areas. All this as well as details, such as whether, for example, test-takers can be expected to use the second conditional in negotiations or produce full sentences in a test of speaking, should be decided before the test.

Mistakes are important in the learning process. It is impossible to learn a skill without making mistakes. This should not be forgotten when testing productive skills. Mistakes show what stage learners have reached in the learning process, how able they are to use and construct rules, how much they are able and willing to take risks, to use all the resources they have available, to be creative and much more.

*Interlanguage* is the term used to refer to the language used by a learner at a particular stage of language learning. Interlanguage is limited and includes features often considered as mistakes. The theory was first proposed by Larry Selinker in 1972 (Interlanguage, in: International Review of Applied Linguistics, 10, pp. 209–241) and describes the differences between the language which learners produce and that of native speakers or highly competent speakers. Interlanguage can vary from context to context.

The significance of mistakes in a test will depend on various things. These will affect their place in the criteria and scores. When deciding on these, consider whether the

penalties for mistakes made in a test resemble those the test-taker would experience in real life. This does not mean that mistakes should not be penalised at all unless they are fatal, but mistake obsession is usually only found among language teachers, and the development of criteria and scores for tests should take this into account.

## 2 Marking and criteria

In general, the person or people responsible for the setting of the test will have decided on the questions, the answers which can be accepted and the criteria for marking them.

**Remember**

One person should not do the entire job of test setting alone!

In many cases, particularly with small-scale and low-stakes tests, the same people may mark the tests. Here the term *mark* is used for the process of deciding whether an answer is right or wrong and applying the criteria. In large-scale tests, the test setter is not usually the person who does the marking. Marking can be done automatically, clerically or by examiners/raters. All of these markers work with some form of criteria, however basic.

Automatic marking is usually carried out by means of a scanner or optical mark reader (OMR) which reads the marks made on the test papers by the test-takers and compares them with an answer key. The test setter is usually responsible for providing the answer key, which of course in this case has to be absolutely reliable.

Large-scale high-stakes exams often use special answer sheets which have to be marked with certain implements in a particular way. If a marking template is used, the process is similar. Answers must be marked on an answer sheet and are then compared with the correct answers using the template. No qualifications are needed to do this.

Clerical marking is done by humans, but unlike examiners or raters, they need not be highly qualified. They must be able to read the language and identify right or acceptable answers, but do not require special training. Clerical marking is used when there is a fairly small and defined range of answers which are acceptable. It is used for half-open questions and tests such as cloze tests. This range of acceptable answers has been decided on by the test setter and, again, must be absolutely reliable. The marker compares the test-takers' answers to a list of possible answers and makes a decision on this basis. It should be possible for clerical markers to make inquiries of the test setter if they are not sure. This clearly should not happen too often. If it does, it means that either the marker is not capable of doing the job or that there is something wrong with the test.

Human marking is most used in tests of written and oral production. In this book we have used the term *examiner* for an oral examiner and *rater* for the person who marks tests of writing. In both cases these human markers work with criteria. They are trained to apply these properly and to ensure that this is done reliably. Examiners in oral exams generally mark simultaneously, i. e. while the exam is taking place.

▶ Chapter 9

In some oral exams, the test-takers are recorded and their performance is marked later by trained examiners either individually or in groups. This may be a more reliable form of marking but may make the test itself less authentic.

▶ Chapter 10

In all these three types of marking, the same core elements are involved. Marking means making a decision on whether to give any points at all, and if points are awarded, how many points to give. The number of points for each answer or criterion may have been predetermined and fed into the programme for the OMR, in which case the examiners or raters fill in marking sheets which are then read automatically. In most small-scale

tests, examiners and raters fill in the number of points themselves and then add them up to get a score.

## Points

The actual number of points awarded for each correct answer or for each criterion is not important to the marker, whether automatic, clerical or human. The term *points* is here used to refer to a number which is then converted into a score. It is the score for the test which holds the meaning and conveys the result. This is decided by the test setter before the exam is taken.

**Example**

For instance in a test of reading comprehension with ten questions, some questions may be awarded 2 points and some 1 point, making a total possible score of 15 (100 %).
An individual test-taker may get a score of 12 out of 15 which would mean a result of 80 %.

In some cases the total number of points to be finally awarded for a test is dictated to the test setter (this is the case at some universities and business schools). The actual number of points awarded has to be converted into this number. More usually the total is expressed as a percentage. Many BE trainers may be able to decide for themselves on the total number of points the test-taker can get and may make a different decision for each test.

**Remember**

Even if you are awarding the same number of points for each answer throughout a test of different skills, you should not compensate for a larger or smaller number of questions by adding the points to a different skill. This will make different versions of the test difficult to compare with each other.

However, even in these cases, it is important that scores on different tests can be compared with each other. This is useful for all sorts of purposes and is much easier to do if they can be converted easily into percentages. It is therefore easiest to have a round number, which can be converted into a percentage without too many fractions being involved. This makes any comparison of test results more reliable.

Only using round numbers, while making tests more reliable, has however the disadvantage that the number of questions will be more or less predetermined. If, for example, you want to have ten points for a test of reading and can only think of nine questions on a text, you will have to either abandon the text, find a second text and set eleven questions on it or somehow think up a tenth question for the first text. Although all these alternatives mean extra work, it is definitely worth it in the long run. If the time for the test allows, the second choice is best as it makes the test more reliable on all accounts.

If you are using only closed questions in a test of receptive skills, you may decide to award the same number of points for each correct answer. In this case, you will not need criteria for the awarding of points and this can be done more or less automatically (even if a human being is doing the actual marking!). If you are using half-closed or open questions, answers may be partly right or partly wrong. You may wish to award half points, fractions of points or use a marking scale. In this case, as well as with most tests of speaking and writing, you will need criteria in order for examiners and raters to make the decision on how many points to award for which answer.

Sometimes so-called bonus points are given for particularly good ideas, long answers or complex language. They may also be awarded to individuals for participation in class or better results in other forms of assessment, such as practice tests. Although this usually has the effect of making the test less objective and therefore less reliable, it can be justified as long as the criteria for the awarding of bonus points are made clear to all concerned. These should also not affect the final result too greatly, e.g. make the difference between a pass and a fail or allow a score of over 100 % to be achieved.

## Criteria

Criteria are necessary for any half-open or open questions as well as for most tests of productive skills. Criteria are necessary to make the decision on whether to award points at all and, if so, how many. Even if an answer given to a half-closed or open question can only be given the full number of points or no points at all, criteria are necessary to make the decision. The number of points to be awarded may not be important to the person marking the test, and so is sometimes given a letter instead. The rater or examiner marks A, B, C, or D, and these letters are converted into numbers later. The test setter however has to decide what the letters mean in terms of numbers. This is what is known as *weighting*. If each criterion can be awarded the same number of full points and the same number of points in each band, then the weighting is evenly distributed.

There are two stages to writing criteria. The first is to decide what your criteria are for the test and give them names. The criteria you use will depend on the test construct, i.e. what you are testing. This is what makes them valid. If you are testing understanding, you may wish to completely ignore grammar mistakes, to take them into account in certain cases or you may want answers to be absolutely correct in every sense to get full points. As long as the construct is evident throughout the test, the test is likely to be valid. Validity applies just as much to the criteria as everything else in the test.

**Remember**

If a test of speaking claims to be at CEFR level B1 and includes communicative tasks where test-takers are expected to speak freely, it may make the test less valid if points are deducted for grammatical mistakes.

With half-closed or open questions where the answer only consists of a few words, the criterion may be simply correctness of information and language, i.e. the answer must be correct in terms of its content and also correctly written (spelling, grammar etc.). This can apply to tests of reading or listening comprehension. The only criterion may be whether the information given by the test-taker is complete and correct or not. If you also want to test correctness of language, you would have to include a second criterion, such as language accuracy (spelling, grammar etc.). These two criteria may be combined, so that both have to be fulfilled to get full points or you may wish to have a scale of points for each answer, for instance from 0 to 3.

Whether you have a scale of points for each answer or not will depend on the type of test, the level and the time you have available for marking. The various points on the scales are often referred to as *bands* and, just like the criteria descriptions, they have to be defined. There may be only two bands – right and wrong – or there may be several, as in most tests of speaking and writing.

In general the criteria for written answers in tests of reading and listening fall into three categories: correctness of information, comprehensibility and correctness of language, although they may have different names. There are usually three or possibly four bands, corresponding to completely correct, understandable but with mistakes, cannot be understood, wrong information or no answer.

With tests where there are no correct answers (tests of speaking, for instance), you will need bands as it is unlikely that in a test of speaking or writing for BE, everything that is said or written can be marked either as right or wrong (full or no points). There will be gradations of correctness in each answer ranging from absolutely correct in terms of the criteria (full points) or absolutely incorrect or unacceptable (0 points). How many bands you have will depend on various things, such as the total number of points to be awarded, what grades are to be awarded or what feedback you want to give the test-takers, among other things. It is best to have an even number of bands and mark a half-way point (usually pass/fail) as if there is a middle band, all examiners and raters tend to overuse it. This may however not apply if you use a bottom band with 0 points. A band with

0 points is often provided for reasons of consistency, even if it is never or extremely rarely used. It can then sometimes be disregarded in the number of bands.

The question of whether to allow the possibility of 0 points is often raised. One view is that 0 points should be given only for no answer at all, and that any language which has been produced should get at least one point, even if this means nothing in terms of the final result. This is to encourage language production rather than simply no language at all as well as intelligent guessing, and definitely has advantages in terms of feedback and communicative language use. The other view is that the answer must correspond to the task in terms of content, regardless of the language, in order to get any points at all. This can mean that a linguistically completely correct answer would get 0 points if it bears no relation to the question or task. This approach places the emphasis clearly on communication and understanding rather than simply producing correct language. How you decide to deal with the issue of 0 points will depend on your view of language and your test construct. The question is relevant to all tests where language of any sort is produced by test-takers.

▶ Chapter 2
▶ Chapter 15

Another question is whether minus points should be given for wrong answers. The only justification for this is that it discourages guessing, and the only place where it may be appropriate is in a placement test. It is generally not done in language tests and is only fair if the marking and the reasons behind it are explained clearly to test-takers. There is no real justification for deducting points for incorrect answers, in view of the role of intelligent guessing in language production and of guessing as a useful test-taking technique. Deducting points can also mean that a test-taker can end up with a negative score, even if a great deal of language has been produced.

If all answers given clearly fall into one or other category, clerical marking is (theoretically) possible and you will only need examples of right answers. However if the answers given are less clear cut and a wider range of answers is possible, you will need to provide descriptions to cover all eventualities, including 0 points. Markers using the descriptions will have to be trained in some way, or at least familiarised with the descriptions, know what they mean and be able to use them. The descriptions are also used to justify scores given to test-takers. They usually form part of the specifications.

**Remember**

Pre-testing is really a very important part of test setting. You should always try and organise some sort of pre-testing, even if it is only on a very small scale.

It is always useful to provide sample answers for each score. The test setter should try and answer the questions or complete the tasks as if taking the test to see if it is possible. If pre-testing is done, the answers given can be used to provide examples of the different scores which can be awarded. Sample answers are most useful if they are obtained from learners during pre-testing. The criteria must be written so that all answers given can be accounted for.

▶ Chapter 4

In tests of speaking and writing, there will generally be more than three criteria, although in tests of speaking not usually more than six at the most. It is difficult for examiners marking in real time to deal with too many different criteria at once, unless the test is very long and is divided into several parts, with each part marked using different criteria.

In general, the criteria for tests of speaking are interaction, fluency, pronunciation, vocabulary range and linguistic accuracy. Most criteria for tests of speaking fall into these categories, although they may have other names, such as task management, speed of speaking, phonological control/intonation, grammar and vocabulary (range and accuracy).

It is important that you define criteria clearly and that they do not overlap with each other. Test-takers should not be penalised or rewarded twice. Everything said above about

descriptions (for bands) and sample performances applies just as much to tests of speaking and writing as to other tests. In tests of speaking, the descriptions should be internalised by oral examiners before the test as well as available for them to refer to during the test. They may be abridged for use during the test itself, but they are essential for reliable and valid marking, whether this is done by one or two examiners.

**Remember**

Two examiners or raters are always better than one!

▶ Chapter 10

## How to write criteria

The main aim of criteria is to give markers, examiners and raters a basis for their decisions (in the following the term *marker* is used for all these three groups). Criteria are necessary to make marking and scoring as reliable and objective as possible. Criteria must be clearly understandable and easy to use. They should be formulated in language which all those who are using them will understand as they will also justify decisions to test-takers. This is why there are sometimes two versions of criteria, one for markers (BE trainers) and one for learners. The content should be the same, but the language used or the length of the descriptions may be different.

Criteria may also include marking instructions so that they are easy to use. This may mean giving instructions on what part of a test to mark first. For example in a test of extended writing, it is sometimes helpful to give the marks for the entire message and structure of the text first and then go on to the marking of the language. Depending on how experienced the markers are, instructions like this can be very helpful. They are also useful if a group of people is marking the tests, especially if they are not doing the marking at the same time.

Criteria describe what is being marked, e. g. accuracy of language. The rating scales give the bands according to which this criterion is marked and the scores for each band. As mentioned above, in some cases the exact scores are not given, but raters just decide which band the answer belongs to. However if scores are given and markers add these up themselves, it must be clear whether they can in fact only award the exact numbers allocated to each band or if numbers in between or half points are possible.

**Remember**

English teachers and testers often make mistakes adding up points! Any calculations should be as simple as possible.

Like criteria, bands and rating scales must be clear and easy to work with. In some cases, there may be longer descriptions and justifications for criteria, bands and rating scales available, but the ones used in the actually marking should be precise and to the point. For tests of speaking they should really fit on one side of a page so that they can be referred to easily during the test. In tests of writing they can be longer, but should not be complicated.


*"The scores for our latest standardized tests are disappointing. We have to articulate a coherent policy of obfuscation."*

**Remember**

It rare that any test-taker will make absolutely no mistakes at all in a test of speaking.

The wording is also important. Try and make criteria and band descriptions parallel as far as possible. The top band and the lowest may be much simpler. For the top band, for instance, it could read: *The task is dealt with completely with no or very few mistakes and none of these affecting understanding* and for the bottom band: *No answer/wrong answer/answer incomprehensible*. Any bands in between the top and the bottom should make the differences between them clear. For this reason it is best to avoid words like *good, enough, adequate* as well as adverbs like *sometimes* and *usually* as these are often interpreted differently by different markers. Make it clear whether the criteria are inclusive or alternative, e.g. *The information given in the answer is correct and there are no language mistakes* or: *The information given in the answer is correct, but there may be several mistakes, none of which affect understanding*. Remember that if you use a word like *no* in the sense of *no mistakes*, this must mean what it says. For this reason, the expression *no or very few mistakes* is often used for the top band in tests of speaking.

Some tests include a criterion and score for general impression. This is however often difficult to define and can be highly subjective and therefore not very reliable. If only one or two people are marking the tests and they can agree on what it means, it can be used to prevent the marking of the test becoming too analytic. Some BE trainers feel happier giving a mark for general impression as it corresponds more closely to what they do in teaching. However in general, what is meant by *general impression* in speaking is covered by the criteria *task management* or *interaction* and in writing by *message* or *structure*.

### Transparency

Criteria and scores should not be a secret. If they are well defined and easy to understand, there is no reason why they should not be made available to test-takers. It is usually only when they get low scores and poor results that learners really become interested in why this happened. It is then that criteria can be very helpful in making clear that both the test and the result were fair. It is always worth considering whether and exactly how you can justify a low score in a test and this is exactly the purpose of objective criteria. The criteria may also be useful if you give written feedback after a test. If the same wording is used in the feedback for different test-takers they will see that they are being treated the same, i.e. fairly.

▶ Chapter 15

## 3 Scoring and Grading

### Scoring

The weighting will depend on the type of test and test construct. Equal weighting may be given to all questions and all parts of the test, for example, 1 point for each answer and 10 points for each skill. This makes marking and scoring easy and gives all test-takers an equal chance. However there is often a case for weighting different questions differently, depending on how easy or difficult it is felt they are or how long and complex the answers given have to be. The weighting may also depend on how important the part of the test is in a particular BE context. If several skills are being tested in one test, they may be weighted differently, with more points, for instance, for speaking than for reading or for listening than writing. There are no hard and fast rules for this as the decision will always depend on the circumstances.

**Remember**

The marker, examiner or rater does not necessarily need to know the number of points for each answer or band.

If only one skill is being tested, it is very important that the weighting reflects the aspects of this skill. If in a test of reading comprehension test-takers are asked to write short answers to questions, then they should really be able to get a good result if the answers

contain the right information, even if they contain language mistakes. This does not mean that language mistakes should not be penalised, but that the result should reflect how high the test-takers' level of reading comprehension is and not how good their grammar is. Similarly with a test of speaking, you may wish to give fewer points for grammar and vocabulary and more for interaction. The weighting will reflect the test construct and the meaning of the results. It should also be explained to test-takers so that they know what to concentrate on in the test.

▶ Chapter 10

The same applies to the decision on the number of points. If you award a large number of points to each answer (large in this case means more than 1) and do not have many questions, this will mean that one wrong answer will be far more significant for the results. Smaller steps between the number of points for an answer or the scores in bands mean a smaller margin of error. This is sometimes compensated for by an instruction that the test-taker should always be given the benefit of the doubt, but if the benefit of the doubt is given too often, a possibly weak or borderline test-taker will end up with more points than is really justified by their performance in the test. There is a greater chance of an error in the test result. This can be avoided by using more questions and smaller numbers of points as well as smaller steps between bands, but this makes setting the test and defining the criteria more complicated and time-consuming.

If your learners are taking an external exam, it is important that they know how the different parts and criteria are weighted, so that they can best prepare for the exam. They may also use the information to make a decision about the order in which they do the parts of the exam.

▶ Chapter 6
▶ Chapter 15

## Grades

The term *grades* is used here to mean the final result and how it is conveyed to the test-takers. The result of a test usually means whether the test-taker has passed or failed the test, and the grade is how well or badly this was achieved (or not). The result may be conveyed to the test-taker in the form of a percentage, but is in many cases, apart from placement tests, also in the form of a pass or fail and a grade such as 1–4 or *excellent, very good*, etc. This means making a decision on the cutscore. This is the point at which the test-taker has reached the level or passed the test and will be logically somewhere between 0 and 100%. With placement tests, the decision made by the test setter or institution is a similar one. One placement test may have several cutscores or cutoff points.

▶ Chapter 2

The pass/fail score or the points given for grades are often decided by someone other than the test setter. In a university context, the score required to pass may be 50% or 60%. Although this is the final decision made when marking and is not usually decided by markers themselves, it is vitally important that the markers', examiners' and raters' decisions correspond with this. It should therefore be clearly reflected in the criteria and the weighting. Depending on the weighting, it may mean that test-takers can pass an exam without doing one part at all if they get above a certain number of points in all the other parts. There is nothing wrong with this if it is intentional on the part of the test-setter and the institution.

Although the pass/fail score and the grades affect all the other parts of scoring and marking, they may not be important in all BE contexts. If you are setting and marking the test yourself, you may not have a pass/fail score or grades at all. However it can be of great importance in high-stakes exams such as at universities or in companies where promotion may depend on getting a certain score and grade in a test. Many test-takers

are used to a system of grades and appreciate it, so it is worth thinking about even if you feel that is too reminiscent of school to be appropriate in a BE context. Grades as well as detailed scores can be useful for giving feedback as well as to motivate learners. If you are preparing learners for an external exam, they will definitely be interested in what grade they get. This information may also be important to (potential) employers.

## 4 Examiners

Language tests should generally be marked by people competent in the language concerned. This however is not enough on its own. Markers should also be entirely familiar with the marking criteria and know how to apply them in practice. The best possible way of ensuring this is to have a session where potential markers mark the same papers and compare and discuss the marks they give, using the criteria. If this is done at an early enough stage, insights gained can also be used to improve the criteria. Ideally criteria should always be tried out to mark real tests to see if they work. In small-scale tests, criteria are usually tried out on "live" test-takers and then possibly adjusted for the next test. This is not ideal, but as long as the criteria are not changed to the disadvantage of the test-takers, it is often more practicable. What is not fair and should be avoided (except possibly for the benefit of the test-takers) is to change the criteria during the marking.

**Remember**

Test-takers should not have to suffer because of bad tests or faulty criteria!

Even individual markers, marking all the tests themselves, do not always mark consistently, which is why exam boards carry out studies of intra-examiner/rater reliability. How you mark tests will depend on your mood, the time of day, how tired you are, how many you have already marked, what sort of language you have been exposed to and what the answers given in the previous test paper were like. If you are using criteria, remember that it is these which are the basis for your marking and nothing else. Test-takers should not be compared with each other but only with the criteria.

If two or more examiners do not agree on a score for a particular criterion or for a test, there should be rules which they can follow so that a final score is awarded. Ideally, referring to the criteria should resolve the difference. If this is not possible, it may be that the criteria need revising. However even with well-designed criteria, there may still be cases where examiners or raters do not agree. The rules for dealing with such cases may be very simple, e. g. the more senior examiner has the final word. There may be practical instructions, e. g. test-takers may be given the benefit of the doubt, or a compromise may be reached which both examiners/raters can live with.

If the criteria have been well-developed and examiners and raters are familiar with them and use them properly, there should be no errors in marking. The most common error made by examiners and raters is not referring to the criteria and/or not using them correctly. Many markers base their marking on subjective judgements or on other criteria rather than the ones for the test. The second most common error is not calculating scores correctly. This can be dealt with by a system of second marking and counter checking.

■ Developing criteria and deciding on scores for a test is both a question of test construct and one of working out a system which makes the test valid and reliable as well as fair to all concerned. It is often a matter of converting a test construct into numbers and is perhaps the most interesting as well as one of the most important elements of test-setting. It brings the test-setting cycle full circle returning to the basic question of what makes a good test. If a test is set well, the development of the criteria for marking may mean revising some parts of the test. This should not be seen as a failure of those parts but as part of the quality assurance process and essential for the test to fulfil its purpose. ■

### To think about:

- Who will be marking the tests?
- Are different parts of the test weighted differently?
- Do you need criteria?
- Does your scoring and grading system justify all the scores you give?

CD-ROM

*"The one inescapable guideline is that test writing should not be a solitary activity."*

Weir, Cyril (1993): Understanding and Developing Language Tests. Prentice Hall, p. 19

# 14 The stages of test setting

## What goes on when a test is being set?

- What do you actually do when you set a test?
- What happens after the test?
- What can you use the results for?

**What is it?**

Test setting is the process of designing the test which includes writing the specifications, items and instructions. In this book it is mainly writing the questions which is meant.

**Remember**

Although it may be the practice of test setting you are largely interested in, the theory can be extremely helpful.

There are many practical issues which have to be considered when you are setting a test. Practicability for all involved is one of the features of a workable test. This includes the work of test setters which should also be made as practical and as practicable as possible. This means setting up a system which can be used again, defining the different tasks that the people involved will have to carry out and recycling good test items, amongst other things. It also means thinking about what goes on during the test and what possible other good uses the test material and results can be put to, so that all the work done is not only done for one purpose only. A test should not represent the end of a process and be filed away never to be looked at again. If the whole testing system is to be used to its best advantage, the test itself can be improved and re-used and the system of test production and administration improved and used again to produce more and better quality tests. Administering a test does not necessarily require any formal qualifications as it may only consist of invigilating or giving results to test-takers.

When you are designing a testing process and setting a test, although you will probably not be able to consider everything all the time, it is best to leave as little as possible to chance. Even if you feel that you have taken everything into account, there will always be the individual test-takers and their own specific needs to contend with, whether before, during or after the test itself. As well as this there will be the needs of all the other

stakeholders, such as the institution or company, parents or potential employers as well as others. These will exist alongside the specific issues relating to the type of test and what is being tested.

**Remember**

There will probably always be one thing you haven't thought of, when setting, administering and marking a test!

## 1 Setting up a system

All those involved should be familiar with the general principles underlying good tests and bear these in mind. It is always helpful to go over these again as even experienced test setters check and re-check their tests to make sure they meet the requirements of good tests. This is the first thing to consider when evaluating a draft of a test: all those involved in the test-setting process must agree with the principles and be prepared to stick to any standards which have been set. This is particularly important if several people are setting tests which are intended to provide comparable results, for instance to be taken by different groups in a company or university.

▶ Chapter 1

Although it is still very common procedure to set a test the evening before learners have to take it, this is something which should definitely be avoided. There are almost bound to be (at the best) typing mistakes in it, at worst far more serious mistakes. Mistakes of any sort in a test will confuse and irritate test-takers and so affect their results in ways you, the test setter, did not intend. If you have thought about what you are testing and why, make a rough draft of the test you intend to set some time before you need it and use it to start the process of producing a good test.

Set up a system for producing and recycling tests. This is often called *item banking* and described in more detail below. Involve as many people as possible and define their roles. One of the best ways of producing valid and reliable tests is through discussion of formats and drafts. If you can produce several different versions of a test yourself or get them from others, these can be compared and the best chosen from a range.

▶ Chapter 4

Decide on the format of the test before setting a particular test. Make sure you can describe this in as much detail as possible. Include information on the construct, but also on the practical details. If you are in charge of a group of trainers who are working as test setters, give them instructions on the parameters of the test. This means defining and stipulating things such as the length of the texts to be used, the format and number of questions, and may also include lists of vocabulary and grammar. If you already have a format, give test setters a sample test. If marking criteria are already available, they are very helpful when writing new tasks. Depending on your situation, you can insist that test setters stick to these. Suggestions for changes to test format should be discussed before the test actually starts to be set. Ask the test setters to make sure that there are no formal mistakes or typing errors in the test before submitting drafts, but also make sure that all tests are proof-read at the end.

**Remember**

One way of proof-reading a test is to do it yourself as if you were a test-taker.

Do not assume that the first draft will be perfect. Even very experienced test setters need other people to evaluate their tests and rarely produce a first draft which can be used exactly as it is. Testing institutions usually give tests to experts for evaluation. These experts may come from different areas of teaching, training or research, and need not necessarily be testing experts as such. In the case of BE tests, test evaluators may well be fairly ordinary BE trainers. If, for whatever reason, there are not many people you can ask to evaluate a test and there are only one or two people who can do it, it is best if they know something about testing principles. As well as testing experts, however, trainers and even learners can be involved in the evaluation process. What should be made clear

to whoever is involved is that you are not interested in blanket praise. Although it is gratifying to hear that the texts are interesting and that the activities sound fun, it does not help much in deciding whether your test is a good one. Different people are good at looking at different things. Some find mistakes of consistency, others find instances of several answers which may be right and yet others make suggestions for making texts or tasks more authentic or easier to understand.

If you have several different drafts of a test, you may choose to use the one which needs least editing first and keep the rest for later. Many tests can be recycled with modifications and used again. Details of items can often be changed without changing the nature of the test or making it invalid, but it is essential to make sure that this is in fact the case. Often changing the order of discrete items or changing the names in texts make the test reusable and not immediately recognisable as the same test.

## 2 Test instructions

**What is it?**

Test instructions are also sometimes referred to as the *rubric*. However the word *rubric* is also used to mean the criteria which can lead to confusion.

▶ Chapter 15

It is often helpful to formulate instructions for the test-takers before starting to set the test or while you are setting it. The test-takers must get all the information they need and be able to understand exactly what they have to do in order for the test to be fair. Examples of good instructions can be found in external exams and standardised tests. There are plenty of examples of bad instructions to be found too.

Make sure that all the words and structures in the test instructions can be understood by the test-takers. There are many words used in teaching and testing which are not part of everyday vocabulary, for instance *gap* and *blank*. Test-takers may also not be familiar with grammatical terms.

**Example**

Examples from tests for low level learners:
*Complete the text below by inserting appropriate words into the gaps.*
*Transform the adjectival into the adverbial form.*
*Unjumble the letters to make proper English words.*
Quite apart from the questionable validity of two of these tests, it is unlikely that learners at this level will know (or even need to know) the words used in the instructions.

It is not a good idea to have to explain test instructions during the test itself. This can lead to anxiety among test-takers which can have an effect on the results and make the test less valid and reliable. Instructions should be short and to the point and any instructions which are recorded for tests of listening should use exactly the same words as on the test paper. In general the same words and terms for the same things (such as *text, question, recording* etc.) should be used throughout the test. Do not refer to the *text* in one place and the *article* in another if you mean the same thing. Also, make sure the order in which the instructions are given is logical. This may be first the task, and then details, then the time allowed, then the score to be given. The logical order will depend on the type of test. If you are giving an example, give it after the instructions. Make sure the example is a good one, not too difficult and that if an answer is given, it is the right one!

**What is it?**

A *sub-test* is a part of a larger test or exam. This could be either a part of a test like *listening comprehension*, e. g. testing a particular skill or one part of this (such as *listening for detail*).

Tell test-takers how many points they can get for each task, item or sub-test. This may affect the order in which test-takers do parts of the test. However in general, test-takers tend to do test questions in the order in which they appear, so think about the order in

which the sub-test and questions are given. Although it makes sense to start with an easy item, important parts of the test should come fairly early on.

In tests of speaking, examiners are often expected to read out instructions. This is to ensure that all examiners use the same words. In good tests the formulation has been carefully considered, so there is usually no reason for examiners to change it. The language of instructions should always be at a slightly lower level than the language level of the test-takers to ensure the highest possible interactivity with the task. In written tests, instructions should always be complete. In a test of speaking, whether or not additional explanations may be given will have to be decided. However if the test-taker cannot understand the instructions, then something is wrong, possibly with the instructions themselves or with the test-taker's level of English.

**Did you know?**

Test-takers regard tests of listening comprehension as more difficult than tests of other skills, so it is not a good idea to start a test with a listening task.

## 3 Face validity

Face validity is what the test looks like. Does it appear to test what it says it is testing? Are those who use it happy with it? It is important that test-takers feel that they are being tested properly and fairly and this of course must be compared with expert judgement of the test. The test must look like a test and must look as if it is testing what we are claiming to test. This is related to content validity, i. e. a test must appear to be testing BE even if we know that the skill of, for instance, listening can just as well be tested using recordings from non-business sources. There are also psychological factors involved. If test-takers' height were measured to test their English they would complain, even if height were a good indicator or predictor of ability or skills in English. A seemingly abstract or theoretical correlation will not be enough in most cases to justify a test where stakeholders do not understand why it has been set in this way.

## 4 Test administration

Of course the process does not end when the final version of a test has been set and printed. In many cases and perhaps in particular in BE contexts, whether in companies or at universities, trainers themselves are involved in the administration of the test. Even if you do not actually invigilate the test yourself, you may have to define how this is done, for example whether dictionaries are allowed, where the answers should be written and whether test-takers can leave the room when they have finished the test, even if the time for completing the test is not up.

**What is it?**

*Invigilate* is a British term used for sitting in on a test to make sure that it runs smoothly. The word *proctor* is often used in the US.

If you do not have to invigilate the test yourself, you may have to produce instructions for this or your institution may have general instructions for the running of tests. This is the case in many universities and colleges.

If you have sub-tests, you may wish to stipulate that test-takers hand in their answers to one sub-test before continuing to the next so that they cannot correct their answers later, for instance after a break. Possibly these things are decided by the institution you work for or may depend on the number of test-takers taking the test at one time.

In a test of speaking, test administration will probably involve making a test schedule, forming pairs or groups and allocating rooms for preparation if necessary. You will also have to consider how important it is for test-takers not to have contact with each other once some have done the test and others are still waiting to do it. Obviously there is no

way of preventing this altogether, but how much contact there is between test-takers will affect the number of different questions or versions you have for one test session and whether there is any attempt to keep the two groups of test-takers apart. It may increase the face validity of the test if test-takers see that matters like this are taken seriously.

▶ Chapter 10

## 5 Cheating and plagiarism

There is a great deal of literature on cheating and plagiarism, how it is defined, why it happens at all and how it can be prevented. Make sure you have defined for yourself what you mean by cheating. It can mean using unauthorised aids, such as dictionaries or notes which have been brought into the test, and it can mean copying or allowing someone to copy during the test. Plagiarism will be more of a problem in longer assessed work, such as term papers or dissertations. It is usually helpful to consult material on this published by exam boards and educational institutions on your particular situation. There are also quite different attitudes to cheating, often depending on culture and age as well as the educational background of the test-takers (and others involved, including invigilators). There are no real hard-and-fast rules on the matter, but as with so many other aspects of tests, it is important to think about what you mean by unauthorised help, cheating, copying and plagiarism.


*"What did you put down for the definition of plagiarism?"*

In general, there is little point in cheating in tests of performance and, for instance, finding out the tasks in a test of speaking before the test will probably have little effect on test-takers' performance and not necessarily improve the results of weak test-takers. In a high-stakes test, however, it is important to make sure that learners do not have access to test questions before the test, even if this is again mainly for reasons of face validity. Reading the instructions for test administrators concerning the security of test papers issued by large exam boards such as Cambridge ESOL will give you some idea of what sort of precautions can be taken to make sure that there is no access to the questions before the test. How many of these precautions you decide to take will again depend on your individual situation.

## 6 Benchmarking and Calibration

The terms *benchmarking* and *calibration* are often used in different ways. Here they are used to describe stages in the standard-setting process and relate to the results of tests or pre-tests. They are closely related to the reliability of a test.

A *benchmark* can be the best result you expect from the test. In a written test, this will probably be 100%. In a test of speaking or writing it should not, except at very high levels, mean a perfect performance. The benchmark(s) will have been decided either using work learners have produced or results of pre-tests. Benchmarking can also be used to describe the process of grouping results into bands so that performances are not judged against each other but against an external set of descriptors such as the CEFR or criteria based on these or other standards. Benchmarking is particularly important if a group of examiners or raters are marking tests, the results of which will be compared with each other. It is useful to mark one or two papers together first and make sure that the same standards are being applied by all concerned.

**Did you know?**

Examiners and raters who often work together tend to mark in more and more similar ways to each other over time. Their marking may be in line with the standards or criteria, but it is also equally likely that it deviates from these, albeit in a similar way! Just because you always agree with one particular examiner on marks does not mean you are both marking well.

The term *calibration* is used to mean an assessment of an item's difficulty to ensure that all items in a test are at the same level of difficulty for the group of test-takers. They may be calibrated for this group of test-takers or against an external scale such as the CEFR. Calibration can be done using expert judgement or statistics, as described below.

The term calibration is also used for the process of bringing examiners (usually in tests of speaking) "back in line with each other". Even when strict standards have been set for test setters and examiners/raters, after a time there is a natural human tendency to deviate from these, either by setting more and more difficult tests or by marking in particular, often idiosyncratic, ways. This deviation, which makes a test unreliable and invalid, can arise from becoming too familiar with the criteria and having too much confidence in your own experience and judgement. Regular examiners' meetings and calibration sessions help to retune judgements and bring them back into line with each other and, more importantly, with the criteria and standards. If you are working in a team, it is important to build these sessions into your testing process.

**Remember**

Benchmarking and calibration are all part of the process of making sure that the test is not "the end of the line".

## 7 Statistics

Test statistics are used to show various things. The most important of these are how many test-takers get the item right, how easy or difficult the item is and how well items differentiate between weak and strong test-takers. There are many more things which can be done with test items, but it would be beyond the scope of this book (and probably beyond the interests and resources of most BE trainers) to deal with them here. However statistics are just as important as expert judgements and evaluations and, if used properly, can provide extremely valuable insights into the validity and reliability of a test.

Test statistics programmes will tell you how easy or difficult items are, how well the items discriminate between weak and strong test-takers, how well distractors in multiple-choice and other items function as well as several other things. However it is generally these three functions which are the most important, especially on a small scale, and in this case it is quite possible to work out these things yourself. You should do some statistical work if you are planning to use tests again, or if the overall results of the test do not reflect what you expected from it. You will obviously be able to form a superficial judgement of the ease or difficulty of individual items, but it is important to back this up with hard facts. It is precisely these that statistics provide. It is important to know how

easy or difficult items are to ensure that the test is at the right level. This is often given as a simple percentage (e.g. 90% of test-takers got the item right).

How well an item discriminates is also important. An item may be easy in general but discriminate well, i.e. only the strong test-takers get it right. If it does not discriminate well, then those test-takers who get it right will be spread across both weak and strong test-takers. If it discriminates negatively, then strong test-takers will be getting it wrong while weak ones get it right! This is obviously undesirable. In multiple-choice questions, it is usually important that the distractors are equally plausible and that all the distractors are chosen equally often. If no test-takers choose one distractor, it should be replaced. The same applies if all test-takers who get the item wrong choose only one distractor.

The results of a standard testing programme look like this:

| Item Statistics | | | | | Alternative Statistics | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Seq. No. | Scale-Item | Prop. Correct | Disc. Index | Point Biser. | Alt. | Prop. Total | Endorsing Low | Endorsing High | Point Biser. | Key |
| 8 | 2–1 | .38 | .52 | .48 | A | .00 | .00 | .00 | | |
| | | | | | B | .38 | .13 | .66 | .48 | * |
| | | | | | C | .12 | .11 | .12 | –.01 | |
| | | | | | D | .49 | .74 | .23 | –.44 | |
| | | | | | Other | .01 | .00 | .00 | –.11 | |
| 9 | 2–2 | .71 | .42 | .42 | A | .07 | .11 | .01 | –.16 | |
| | | | | | B | .11 | .18 | .04 | –.22 | |
| | | | | | C | .10 | .16 | .00 | –.22 | |
| | | | | | D | .71 | .53 | .95 | .42 | * |
| | | | | | Other | .01 | .00 | .00 | –.13 | |
| 10 | 2–3 | .68 | .56 | .56 | A | .68 | .39 | .96 | .56 | * |
| | | | | | B | .21 | .36 | .04 | –.37 | |
| | | | | | C | .03 | .08 | .00 | –.24 | |
| | | | | | D | .07 | .14 | .00 | –.22 | |
| | | | | | Other | .01 | .00 | .00 | –.13 | |
| 11 | 2–4 | .57 | .49 | .49 | A | .18 | .28 | .08 | –.27 | |
| | | | | | B | .15 | .19 | .09 | –.12 | |
| | | | | | C | .08 | .16 | .01 | –.31 | |
| | | | | | D | .57 | .33 | .83 | .49 | * |
| | | | | | Other | .01 | .00 | .00 | –.13 | |
| 12 | 2–5 | .61 | .63 | .54 | A | .09 | .18 | .00 | –.22 | |
| | | | | | B | .20 | .28 | .03 | –.27 | |
| | | | | | C | .61 | .32 | .96 | .54 | * |
| | | | | | D | .09 | .18 | .01 | –.28 | |
| | | | | | Other | .02 | .00 | .00 | –.09 | |

Only the most important aspects of the statistics will be described simply here. The first two columns give the numbers of the items and their place in the test. The third column (*prop. correct*) gives the proportion of test-takers who answered the item correctly. The fourth (*disc.*) shows how well the items discriminate between weaker and stronger test-takers. Similar information is given by the point *biserial correlation (Point Biser.)*.

As these are all multiple choice items, the other information relates to how frequently the distractors were chosen by test-takers. The correct option is indicated by a *.

If you want to produce test statistics yourself, you will need to do a little calculation, but it is not difficult. The method described here uses no technology. Mark all the tests or pre-tests, calculate the final score and then divide the test papers into three groups of equal size. These should represent the low scores, the medium scores and the high scores. Count up how many candidates in the high and low groups got the answer right to each item. You can ignore the middle group for this. Then take each item and add the total number of those who got it right in the high scoring group to those who got it right in the low scoring group and divide this by the total number of test-takers from both groups. If over 90% of both groups get the item right, it is probably too easy. If less than 30% of both groups get it wrong, it is probably too difficult. If somewhere between 30% and 60% of the test-takers in the total number of low and high-scoring groups get the answer right, the test is probably at the right level. If over 80% of all the candidates get all the items right, the test is probably too easy, and if less than 30% get all the items right, it is probably too difficult. You may want to have all the items in a test at the same level of difficulty for the group or at different levels. This will depend on the type of tests and the reasons for the test.

If you want to see how well an item discriminates, subtract the number of test-takers in the group with the lowest scores who got the item right from the number of test-takers in the group with the highest scores who got the item right. Then divide this number by the number of test-takers in each group. The answer will be somewhere between +1 and –1. A result of +1 means that all the high-scoring test-takers are getting the item right and all the low-scoring test-takers are getting it wrong. A result of –1 means that all the weak test-takers are getting it right and all the strong ones are getting it wrong. Again this is something to be avoided.

Here is the calculation:
The total group of test-takers is 30. It is divided into 3 groups of 10. The item we are considering was answered correctly by 8 out of 10 in the high-scoring group and 4 out of ten in the low scoring group. The calculation goes like this:

$$\frac{8-4}{10} = .4$$

A result of .3 or more is usually considered adequate as a measure of good discrimination. This means that the item in the example here can be said to discriminate adequately for the group.

Using these simple methods helps you to identify items which may be too easy or too difficult or which do not discriminate well between weak and strong test-takers. However it is very important to remember that you must also look at the item itself. It is not possible to make a good decision on which items to discard and which to keep on the basis of the statistics alone. Poor items can result in acceptable statistics, and items which do not produce ideal statistics may still be included in tests for valid reasons. You may wish to have some very difficult items to identify the highest level among the test-takers (in a placement test, for instance) or you may wish to have easy items at the beginning of a test, to make test-takers feel at ease when they start the test. There may also be items in a test which statistics show are difficult for the group of test-takers involved, but which are nonetheless essential for making reliable statements about their level. If all the test-takers get particular items wrong or score badly in certain tests or

sub-tests, this may also provide you with valuable information about the teaching they have received. You may also wish to have some evidence of ease or difficulty of items so you can use them in other tests at lower or higher levels. You may wish to use some items for which you have good statistical results as well as the corresponding expert judgement in a range of tests. These items are sometimes known as *anchor items*. As the results of these particular items are known, they can give you useful information about the level of new groups of test-takers.

## 8 Item banking

**Remember**

It is always useful to produce more items than you need and edit and pre-test them.

Item banking has been referred to implicitly a number of times. It is the process of collecting parts of tests to use in future tests. The term *item* can refer to individual (discrete) items or to texts or tasks. All of these may have been discarded during the process of test setting, editing and pre-testing for various reasons. They may have been too difficult or too easy or they may have simply been additional parts of a test which it was not necessary to use for that particular test.

If these tests and parts of tests are collected systematically, they can be used to create new tests with the same format, to fill gaps in new tests and to use for practice material. Items are not only items but are also collections of whole tests which have been used and can be used again. Even if they are not used again, the tests and results can be useful when producing new tests. It is therefore worth devising a system to store them and use the system consistently. Although item banks are often used by large testing bodies, examining boards and academic institutions for research purposes, most BE trainers will probably have several opportunities to use tests again. Once you are fairly sure that all those who took the test, whether employees in a company or students at a university, cannot or will not pass on information about it, you can re-use or recycle previous test formats, test versions or actual tests. If they are from your own tests, you can use them in other training situations. How much you change in good tests will depend on you. It may not be necessary to change anything, but you may wish to change superficial details which do not affect the quality or results of the test.

Item banking does not only consist of saving the test items, but also of saving all the information about the test so that you can compare different versions. This includes information about the groups of test-takers and the results of the test. You can trace the way a test has developed over a period of time and get useful information about things like the level of the test-takers and the changes in this. You can also combine items which show the properties you wish them to have in terms of difficulty or topic or task.

## 9 Uses of tests

Test results are mostly considered in terms of their usefulness for the test-takers, and it is of course this group which is probably the most important among all the different groups of stake-holders in the testing process. However there are other groups who can also make use of the results and other uses to which test results can be put. Whether these are all good is another issue. A blanket judgement is probably not possible.

A great deal has been said about the validity of tests. Only valid tests can really provide meaningful results for test-takers and stakeholders. Test-takers may use the results of tests to apply for promotion or a different job. They may use them to try and exempt

themselves from an English programme at an educational institution or to gain entry to a course of studies. Whether they are successful will depend not only on themselves but also the other party. This is not always something the test setter can influence. Tests results can be used to gain entry to a country, or to obtain a residence permit or citizenship. If tests and test results are used for purposes other than the test setter intended, there are often ethical issues involved. In some cases, you may be able to stipulate the purposes to which the results are put; in others you may be totally unaware of them. The ethics involved in testing and the use of test results is one further issue concerning testing. However useful you may feel language tests are, you may not always approve of the uses to which they are put. Dealing with the ethics of testing is beyond the scope of this book, but there are plenty of opportunities to go into it in more depth. On a small scale BE trainers might want to think about the effects of testing on learners' motivation, whether the results of tests are always positively motivating and why this is the case. You might also think about the view of language many tests are based on (perhaps your own) and whether this is always appropriate in modern business life. These ethical issues may be minor in comparison to immigration issues, but thinking and communicating about them is one way in which BE trainers can contribute to the discussion in the field.

■ Working out a system to deal with the various stages of test setting is the best way to produce a good test. It is a question of combining various things, the different criteria for a good test and the needs and abilities of the test setters and test-takers. Instructions should be complete and concise, and the test should be clear and easy to understand. To ensure that all this effort does not lead to a dead end, the results of the test or the work produced by the test-takers can be put to good use in producing further tests, benchmarking performances and backing up expert judgement with statistical analysis. All these practical details will be worthwhile both in terms of improving the standard of testing at your company or institution as well as improving the professional quality of the BE training being carried out. ■

### To think about:

- How far can you involve others in test setting?
- How can you best avoid cheating and plagiarism?
- What can you use the results of tests for?

CD-ROM

*"Aware of the power of tests, policymakers in many parts of the world continue to use them to manipulate their local educational systems, to control curricula and to impose (or promote) new textbooks and new teaching methods."*

Cheng, Liying / Watanabe, Yoshinori (eds.) (2004): Washback in Language Testing. Lawrence Erlbaum Associates, p. 6

# 15 Other issues

## There is more to testing than just writing questions

- What other forms of assessment are there?
- How can I make sure my test is fair?
- What more do I need to know?

There are a great number of issues connected with testing which cannot be considered only in relation to one type of test or specific skill. Some of these may be side-issues, others are important in a large number of different testing contexts. The individual BE trainer will have to decide which of these are important for the particular context. You may feel that your own specific concerns, such as, for instance, continuous or classroom assessment have not been dealt with adequately in any of the chapters, or you may want to read more about some of the issues, such as fairness, which have been touched on several times. There is no shortage of information on any of these subjects, both in the literature on testing, on the Internet as well as in the archives of the email discussion lists on testing issues.

▶ Further information

### 1 Other forms of assessment

Apart from fairly formal testing there is a range of forms of alternative assessment, including continuous or classroom assessment and portfolio assessment. Continuous assessment which takes place during the course of teaching can be as valid as formal testing, but a great deal of care has to be taken to make sure that this is in fact the case.

There is a danger of giving a highly subjective judgement on an individual's language ability or skills. There may be no problems involved in this if there are no high stakes involved, i. e. if the judgement is not to be used for furthering a learner's professional or educational career. Without many of the features of valid and reliable tests, assessment based on impressions can easily be questioned and criticised by both test-takers and other stakeholders. For this reason, all assessments in high-stakes contexts usually need to be justified. Trainers who have thought about some of the questions relating to formal testing described in this book will find that they are usually in a better position to make what seem to be ad hoc assessments, but which are in fact based on valid criteria such as those which apply to more formal tests. It may therefore be more worthwhile in the long run to set a fairly formal test and use this to validate your own judgement. If the results are the same, both forms of assessment will have been shown to be reliable, and everyone will be happier.

The consequences do not necessarily mean a lot of work. In the case of speaking, for instance, it is usually more satisfying for everyone involved to organise and specify particular occasions when learners' performance is going to be assessed and graded. In general, learners perform better if they know that there are some stakes involved. Creating a situation like this allows learners who are reluctant to speak in class to show what they are capable of doing in English. If you have also made a subjective assessment of the learner's performance throughout the course, you can compare your judgements and even take an average of both the informal and formal assessment you have given.

The drawbacks to subjective assessment are that it is just that. A learner whose teacher says his English is "good" may be sent to London to negotiate a deal and find no-one understands him. Similarly, a student whose English has been assessed as "weak" by her teacher may find she does very well in a course of studies in the USA with other international students. This happens because the judgements are not based on commonly accepted criteria or standards. In this case, if "good" and "weak" were defined in a way that all agreed on, the mismatch in judgements would not occur.

A second form of alternative assessment is portfolio assessment. This means basically collecting samples of work which represent a learner's achievement in English at a certain point in time. These can also be used to show progress over a time period. These are selected by the learner (with the help of a teacher) and may include, in the case of BE learners, emails, reports, business letters, a letter of motivation as well as recordings (video and audio) made by the learner. The use of language portfolios is described in detail by the Council of Europe, and there is a list on its website of all the recognised language portfolios available. The *European Language Portfolios* provide a framework for the documentation of learners' language achievements and progress. They serve a very useful function as far as learning is concerned and have been used successfully with children and also with adult learners in many cases. Whether portfolio assessment will be useful to you in your particular BE context will depend on you, and, far more importantly, on your learners. The most important single aspect of portfolio assessment is that the learner retains responsibility for the portfolio and its contents and keeps it up-to-date. Although this makes it valuable in the eyes of the learner, it may make it less highly-regarded by employers or universities. You and your learners will probably find that a portfolio is most valuable as a learning tool and perhaps as a back-up, justification or complement for either the trainer's assessment of a learner's English or the results of a formal test or exam.

Then there is also self-assessment. BE trainers (as well as employers) often have very personal views on this and do not always believe that it is a valid form of testing or assessment. It is certainly highly subjective and not always reliable. How far you include learners' assessment of their own knowledge, ability and performance in your system of assessment will depend on your attitude to it, your learners and the circumstances you are working in. Self-assessment as a preliminary for placement testing can be very useful and collecting data about learners' experiences of learning in the form of hard facts (how many years they have learnt English, how long ago, where and with what teaching methods) may help you to judge their ability for placement or diagnostic purposes.

Asking learners to place themselves on a scale is another form of self-assessment and the CEFR has a scale particularly for this. By giving concrete examples of performance (can-dos), learners are discouraged from placing themselves too low. They do this for various reasons. They may feel that a course at a lower level will be easier and they will get better results, whereas in fact they will probably only be bored and drop out of the course fairly soon. Other learners may assess their level as higher than it is so they can impress their boss or avoid having to admit that they have lied about their English abilities in their application. There are many pitfalls in using self-assessment, and it may be more efficient to use standard testing methods.

## 2 Fairness

That a test is unfair (for whatever reason and whether justified or unjustified) is probably the most common criticism made of tests by test-takers. Of course, not all tests are unfair, but many of them are, not because teachers want to be unfair to learners, but because they have not thought about many of the things which should be included in specifications before giving the test. Language and perhaps especially BE, given the ever-changing nature of both business and English, is extremely dynamic and it is rare that one single person can foresee all the things which can go wrong in a test of BE.

Tests must serve a purpose. In order for them to give learners the opportunity to do their best, they must be fair. If you remember several of the main points made in this book, you will go a long way to making sure that the tests you set are as fair as possible. There are a number of things to remember.

**Learners are not guinea-pigs.** Try tests out first in class or as a practice before learners take a test for an important reason. Give them all the information they need and answer all their questions. Some teachers say that if learners know everything about the test, they will only learn the things necessary for the test and not for "real life". We go into learning and practising for tests elsewhere, but if learners learn something worthwhile, for whatever reason, then that is a good thing. Just as an aside, it is assumed that what you teach and test is worth learning!

**Justify your tasks, answers, scores and grades.** Learners are often happier if you can give them a reason for a low score, even if they don't like the score itself. It is vital for you to know why you give scores and to be able to communicate this to test-takers. If they know the criteria on which their performance is to be judged and can then see that it has been judged against these criteria, they will know where and why they lost points in a test. Use language they can understand when justifying marks and grades. This will also provide feedback and can be motivating.

**Everyone taking a test should have an equal chance to do their best.** Depending on your test, there is no reason why some or even all test-takers should not get 100%. It is not fair to penalise them in retrospect if it turns out your test was "too easy". It is quite possible that all the test-takers have achieved the goal you were aiming at in an achievement test. In this case, they should get the results they deserve. If you know the test-takers, it is always interesting to compare the results you think each would get with what they actually get. If the scores are far apart, it is worth asking yourself why.

**Don't play tricks on test-takers.** It is not fair to ask them to do things in a test they would never have to do in real life or to mislead them deliberately. To avoid this you can, for instance, tell them what to learn or practise. There is nothing wrong with doing this, even if this is all they learn. Decide whether you will focus on only certain things in the test or make a selection from a wider range. This will depend on the type of test. Then tell learners what will be included in the test (of course not the exact questions!) and stick to it. With learners who find languages difficult it can help if they have a focus. A test is a document of learning, whatever sort of test it is, and the main goal of BE trainers must be to help learners to learn.

**Encourage them to guess in the test** unless, of course, the marking system penalises incorrect answers. Many teachers object to tests where answers can be guessed, such as in the case of multiple choice questions. In fact in real life a lot of guessing goes on when we speak a foreign language and even when we use our native language. After all, if we could only say what we are perfectly sure of, very little communication would take place. Show learners how to guess intelligently because this is what it is all about. In a good test, completely wild guessing should not get test-takers very far. In fact, absolutely wild guessing is extremely rare. Most guessing in tests is intelligent guessing and is based on some knowledge or ability.

*When can I stop making wild guesses and start making intelligent guesses?*

**Let learners learn some things off by heart if they find it useful.** Many teachers object to tests where they feel that learners can learn the "answers" off by heart. This usually applies to performance tests of productive skills, such as speaking and writing, where a test-taker has to introduce himself or to make a presentation or write a letter of application, for instance. Learning off by heart, like guessing, is also something we do in a foreign language. How else would we learn chunks of language and be able to use them appropriately? Introductions and presentations are often rehearsed which is very similar to learning off by heart, and is in fact often exactly what makes them good. The skill in producing language which has been learnt off by heart is in making what you say seem

like real communication and possibly (depending on the level and type of test) being able to deal appropriately with the unexpected. The test itself should be able to distinguish between those who have "only" learnt off by heart and those who have also learnt some things by heart, for instance with the following tasks:

- Giving a presentation and answering questions at the end
- Introducing yourself briefly and answering questions
- Discussing one of a range of topics
- Answering one of a range of letters and including certain points

and in the scoring:

- Giving points for spontaneity and fluency rather than only accuracy
- Giving points for natural intonation

## 3 Bias and discrimination

Many texts and tasks in tests of BE may seem rather bland and anodyne to some trainers, who have been used to dealing with possibly controversial issues in the classroom. Whereas the classroom offers trainers possibilities for dealing with possible negative effects of controversial material or emotionally laden topics, an emotional reaction in a test can affect a test-taker's result and make the test less reliable. Large-scale tests are often criticised for not being interesting, but there are good reasons for this. Topics, texts and tasks which upset test-takers as well as issues which favour some groups (cultural, national, age, gender, professional) over others make tests unfair. Of course it is impossible to know what might upset someone, and so the better you know the group of test-takers, the easier it will be to deal with this issue. The less you know about the test-takers, the more you will have to fall back on general topics which will upset no-one and give no advantages to certain groups of test-takers.

## 4 Washback

*Washback* is the term used for the effect testing has on teaching and learning. It has been referred to implicitly several times, for instance in the sections on test preparation and in the relationship between teaching and testing. The term *backwash* is also sometimes used. Although basically interchangeable, backwash is often considered negative, whereas washback, used in the context of testing, can be either positive or negative. Looking at the washback of a test is one way to avoid the "end of the line" effect, regardless of the conclusions drawn on the issue. Negative washback can take place when language is defined in only one particular way (e.g. in terms of grammatical knowledge or translation) for testing purposes and so limits what can be done in class when preparing learners to take the test. Positive washback is when what happens in the test has an effect on classroom activity, making it more authentic or communicative. One example of this is the greater use of pair work in class since the introduction of the paired testing format into examinations such as those offered by Cambridge ESOL. Before pair tasks were used in FCE and CPE, pair work was sometimes done in class. However many teachers seemed to feel it was more of a fun activity rather than a part of serious language teaching, let alone exam preparation. As it is used in several large scale and high-stakes tests, teachers all over the world are obliged to use pair and group work in class in order to prepare learners for these. Whether this is a positive or a negative effect of the test may be a matter of opinion.

If you are evaluating a test which you have not written yourself, you should consider whether the question and task types used in the test are so different from those you use in class as to mean that you would have to change your teaching to accommodate them. If you feel you would have to do this, then you are probably experiencing the effects of washback. Whether you do it or not will depend on how important the test is and how firmly entrenched your views on teaching are.

*Washback* and *impact* are different from each other, although both are about the effects of testing. Whereas washback is connected to teaching, impact refers to the effect test scores or results have on the lives of the test-takers, possibly in terms of employment, residency or citizenship. Impact is also clearly an issue in educational settings. On a large scale both washback and impact are extremely important and worth more research, but on small scale they can also play a role in the design and running of sets. BE trainers will have to think about both washback and impact when they set tests, in terms of the effect on teaching in the company or institution as well as in terms of the significance of the test for the test-takers' employment or educational career.

▶ Chapter 1

## 5 Differences between teaching and testing

Professional test setters often believe that teachers cannot set tests and are sometimes unnecessarily highly critical of their attempts to do so. This may be because many teachers and trainers are not interested in testing and therefore approach the whole subject rather light-heartedly. For teachers and trainers, testing is only one aspect of their work and therefore criticism of teachers' tests should not imply any criticism of their teaching. Good teachers are not always good test setters, and there is no reason why they should be. There is, however, also no reason why a good teacher should not be a good test setter. The only thing that both teachers and those who commission them to set tests should bear in mind is that teaching and testing are different. In the same way that someone who speaks two languages well is not always the best person to translate between these two languages, a good teacher is similarly not always a good test setter. Some of the things that teachers are used to doing in class are completely inappropriate in a test, such as helping, practising and having fun. Similarly answering multiple-choice questions and penalising mistakes seriously are two things which should not be happening in class.

## 6 Theory and practice

The focus up to now has been mostly on the practical aspects of testing, those which most closely affect BE trainers and their work. Testing theory is also important and if you approach the whole issue of test setting seriously, you will not be able to avoid finding out more about the theory of testing. If you want to learn more, there are a number of possibilities, starting with the books and websites mentioned in this book and even doing a course in testing.

▶ Further information

You may find you start looking at all sorts of things through tester's eyes. If you start doing this, you will find that there are lots of so-called tests which are in fact nothing of the kind. The word *test* is often used just to make a series of questions seem more serious or academic. You may find you are reading tests in women's magazines (*Do this test to see if he really loves you*) or in consumer magazines (*50 DVD players tested in this issue*) to see

if they correspond with what you know about tests of English. Some other types of tests exhibit some of the features of good language tests, such as a construct or criteria, whereas others cannot really be described as tests at all. This is also the case with tests in other subjects. The vast amount of work done in language testing, in particular in English, and the advances in this field have led to language testing experts being asked to help with the designing and setting of tests for other non-language subjects such as maths or medicine.

Many language tests are based on performance, others on knowledge or ability. There is a whole area of testing which has little to do with these three things, and that is the area of personality testing. It is much more difficult to test aspects of personality than knowledge or performance. Although personality tests are not in general a part of language testing, they are often used in tests of intercultural competence which is becoming an increasingly important part of BE training. There are many tests of this type in course books and while they may be fun, interesting, good for discussion and for practising English in class, they should not be regarded in the same way as language tests. As they are often called "tests" for marketing purposes, are in English and are used by companies, they may seem to be tests of BE. Many, if not most, of the tests on the market have either only weak theoretical foundations or are fairly easy for the test-takers to see through and adjust the answers they give accordingly to get the best results. They are not usually suitable for sorting test-takers into groups which would be useful for BE courses. Many of the problems associated with these tests also occur in self-assessment tests as the questions are difficult to answer honestly as well as easy to answer untruthfully. In general, personality tests are tests without a recognisable construct or criteria. If you are exposed to them, knowing something about language testing can help to recognise their shortcomings.

■ Although you will probably start with the practical side of testing, if you find it interesting, you will undoubtedly progress to a more theoretical level than can be provided in a book like this. Issues such as validity, fairness, washback and impact, bias and discrimination are all fields of their own, and there are many expert opinions and findings available. You may want to rely on these or you may wish to make our own judgements and even do some research of your own. ■

CD-ROM

### To think about:

- Are your tests always fair?
- What other sorts of tests have you taken or been involved with?
- Where do you see the links between teaching and testing?

*"Practice and feedback are the essential ingredients in developing competence. And in its widest sense, feedback highlights both strengths and weaknesses, successes and failures."*

Brieger, Nick (1997): Teaching Business English Handbook. York Associates, p. 110

# 16 Preparing learners for tests and exams

## Teaching or teaching to the test

- How much preparation should learners get before a test?
- How can they best be prepared?
- How much preparation should they do themselves?

Most teachers and BE trainers probably feel that teaching time is best used for just that, and not for specific preparation for a test or external exam. It is certainly the case that the best tests are ones which do not require a large amount of specific preparation, but it is also the case that a certain amount of specific test-related preparation will help learners to do their best in any test they may have to take. How learners see this will depend on how important it is for them to pass the test and how well-equipped they feel for it. If you are setting a test for your own learners, you can help them by giving them as much information as possible about it (without of course revealing the exact test questions). This is the point of specifications and also what most reputable exam boards do. There may be no need to do much more than this. Learners usually expect a great deal more preparation for high-stakes and/or external exams. If learners are getting grades for a test or assessment, the assessment should be identified as such, and they should be informed about as many aspects as possible. It should therefore be separate from teaching and not confused with teaching in terms of when it takes place and what happens. It is not fair to assess learners while you are teaching them (and they are learning rather than performing) without them knowing that the assessment is going on. It is better to make clear to them what the focus of each part of the lesson or course is, i. e. teaching, practising or being tested.

▶ Chapter 4

If learners are being prepared for an external exam, there is a danger that this will dominate teaching and learning. However, in spite of the high stakes involved for learners in an external exam, teaching time should not become purely test preparation time. Preparing for an exam means several different things. There is firstly the language learning itself. This should be no different in essence from what goes on in courses which are not leading to an exam. There is then a focus on those aspects of language ability and use which are to be tested and also specific exam practice with sample material. As well as this you may want to give learners help with test-taking strategies. Every BE trainer teaching a course where a test is to be taken at the end will have to decide how to combine these various things best.

There are ways of maximising the use of teaching time so that the learning aims of the course can be achieved and learners can at the same time be best prepared for a test they may have to do. This applies equally to a test you have set yourself. Clearly what you do in your lessons will depend on various factors as well as tests, but there are a number of general principles which can be observed. These apply to all achievement and proficiency tests, to a lesser extent to other types of test.

Remember that results in tests say something about test-takers' knowledge, ability and competence. This is, after all, their main reason for taking a test. Just as a certificate from an external examining body can be a source of pride to a learners, so can a test which you set or feedback you give in any form at the end of a course motivate and encourage learners to continue. This should be the aim of any tests you set yourself for your learners: not to trick them or show them what they cannot do, but to give them feedback on their progress, achievements, strengths and weaknesses, depending on what type of assessment or which test you are using.

▶ Chapter 2

## 1 Teaching to the test and authenticity

If you are setting a test yourself, you will probably set it so that your learners can perform at their best. If they are taking an external exam and you cannot influence its content in any way, the situation will be quite different. If a test is good, i.e. from a reputable examining body, you may feel that specific preparation is unnecessary. Why should we need to prepare learners if tests are valid, authentic and interactive and possess all the other features of good tests? Is it not enough to teach them and train them for the authentic experiences of language use they will encounter in real life and assume that these will be reflected in the test tasks? However good the test may seem to be, there are still many good reasons for preparing learners, not least of all that they only have a limited number of chances to show how good they are. In this sense, a test is like a performance which needs to be rehearsed. In the same way as learners want to be prepared in real life to be able to give a good presentation or perform well in a negotiation, they want to be prepared for a test. The stakes are high, and it is only fair that they should get some idea of what is coming and be able to deal with it.

However, a good test should not require an enormous amount of preparation provided the learners are at the right level and have some experience of the type of tasks they will encounter. If the tasks in the test are authentic and interactive, test-takers will understand what is expected of them easily. If there are examples of similar tasks in the course material, all that may be necessary is to point this out to learners every time you do one of these (for example, multiple-choice, true/false, discussion, information gap). In general course activities will consist of far more than just these tasks and so it will

probably also be necessary to point out to learners why it is useful to do other activities and how these will also help them (even if "only" indirectly) in the test.

It should be made clear to learners that, although only certain things will be required of them in a test, they are firstly not learning just for the test and secondly, all the other things they learn will help them to get better results. This applies in particular to enabling skills and learning strategies. Learners will need to know and be able to use grammar and vocabulary appropriately in order to deal with reading, listening, speaking and writing tasks in a test. They will also need to be able to deduce the meaning of words and phrases, compensate adequately for anything they do not know or forget, use appropriate interactive strategies, such as turn-taking strategies, in tests of speaking as well as deal with a possibly nerve-racking situation.

It is to be hoped that you are not teaching only so that your learners can pass a test. If you are really only preparing learners to pass a test and not necessarily to use BE in any way, then simply repeating examples of exam tasks may help them. However even these will have been taken from a pool of language, which will probably be impossible to cover in this way. Without more input and practice, practice tests will not be enough.

If the test or exam includes items and tasks with which learners are not familiar, then you will need to spend some time on these to make sure they know what to do. If, for example, in the exam the test-takers need to write a business letter and have only ever written essays in the course, then you will need a great deal more time to teach them this than if they are familiar with the task already. It will not be enough to just give them examples of exam tasks and ask them to do these. If writing letters is something where they need to have proof of competence, there will be no way round teaching them what they need to know and finding material for this. However if they are looking for a general qualification and do not need to write business letters, it will make more sense to look for an exam where this is not required. The same applies to other tasks. There are exams, for instance, which test only oral competence. This may be all the learner needs.

▶ Chapter 6

This may mean using specially produced exam preparation material, which many exam boards provide. This is not the same as using practice tests. Preparation material, if it is good, helps learners to concentrate their energy on giving the right answers rather than working out how to do the task. Examples of exam tasks may be used in preparation material, but it should consist of more than these.

Tasks and activities you do in class are almost always more authentic, i. e. more related to what learners do in real life, than test tasks. Of course this varies, but it is generally the case. Exams are usually written for large groups and according to certain standards and so cater for a broader range of needs. The smaller the group of test-takers or learners, the more the task can be adjusted to their needs and abilities.


This does not mean that classroom tasks are not helpful, on the contrary. It is how you use them as test preparation that is important. Even if you cannot mark them as if they were in a test,

you can give learners useful feedback. Concentrate on one or two criteria for each learner and help individuals to understand what they can already do well and what they still need to focus on before the exam. It probably goes without saying that you make the usefulness in real life of classroom tasks and activities clear to learners, so this means simply transferring this to the exam they are going to take. The more authentic the test is, the easier this will be. Point out to learners which aspects of classroom tasks will be useful in the test. This will not only give them the help they need, but also make them realise that what they are doing in class relates to the test and to real life.

▶ Chapter 4

## 2 Test or exam – low or high stakes

Learners will also most probably expect to know and even to be shown what the test looks like. A description may be enough, but examples are usually more helpful. If the test you are preparing them for is one you are setting yourself, this may mean that you have to provide practice tests (mock exams) yourself. This may seem like unnecessary extra work, but in fact it serves several useful purposes. Setting a mock or practice exam can help you to identify the areas of testing you need to focus on as well as any weak points in the test. Doing this before the real test will help to make the test better. If you are setting a test yourself, it is only fair to inform learners about what will be in it. Give them as much information as they want and as much as possible. If you have to set an end-of-course or end-of-term test, you should really think about what sort of test this is to be and all the issues involved before you start designing your course or collecting your teaching material. Although it may seem hard work to produce practice tests, they can be of great help not only to your learners but also to you when you set the final test. You can use them to develop the criteria and to benchmark samples of learners' work as well as to see if the test is too easy or too difficult for the learners. If you are using individual items to test grammar and vocabulary, you can often use these with modifications in other tests. If you have to set new tests for each course or each term at a university, you can use past test papers for practice tests. Parts of tests can often be combined, with slight modifications, to make new tests. If the test is fairly low-stakes, there is no reason why you should not be open about some of these issues with learners.

**What is it?**

The term *mock exam* is often used for an exam where the results mean something, i. e. they are combined with the results of the real exam to get the final result. The term *practice test* or *practice exam* means just that. However, the two terms are also often used interchangeably.

Practice tests are not the same as teaching material. They are useful but should definitely not become the focus of your teaching. This is the case whether you set the practice tests yourself or use ones provided by an exam board.

Make it clear to learners what the results of practice tests mean. Learners should not be disappointed if the results of the real test are lower than those of the practice tests. It is usually a good idea to make practice tests slightly more difficult that the real test or mark them more strictly. If practice tests are provided by someone else, you should be able to assume that they are at the same level as the real tests, but will not be able to take any responsibility for this. Make this clear to learners before they take the real exam.

If an exam board provides several examples of exam tasks in the same format as the actual exam, then it is usually enough to do just two, one fairly near the beginning of a preparation course and one just before the actual exam date. Reference can be made to the exam in between these two practice exams and work can be done on certain aspects of it. Do not do the practice tests earlier than at a point where your learners have reached the level they need. Apart from it being boring to do too many practice exams, learners will probably find the tasks too difficult if they are done too early and this will be frustrating. Practice exams are just what they say they are and for this reason should also

not be done too long before the exam date. Allow some time to discuss the results of the first practice test and perhaps concentrate a little on problem areas (and hope there will not be too many!) in order to improve for a second try.

Let your learners do the practice exams under exam conditions, i. e. individually, without any help and in the time allowed. Then mark them using the answer key and criteria given. After the first one, discuss the difficulties the learners had, what they can do about these and decide how you can help them. This will include the tasks themselves, the instructions, the time involved and how happy they are with the results they get. If you focus on what they need, the results of the second practice exam should be better.

The intention of practice exams is to give test-takers an idea of how well they would do in the real exam and help them to decide if they possibly even want to take it. Mock- or practice tests are usually produced under the same conditions as the real exams and have gone through the same test setting and editing processes as well as been pre-tested. If your learners are taking an external exam, it is not a good idea to try and produce practice tests yourself. You may be able to set a test which looks like the external exam your learners want to take but it is highly unlikely that you will have the resources to make it of such high quality. Practice exams are sometimes set at a slightly higher level of difficulty than real exams so that test-takers do not experience nasty surprises in the real exam. They should however be at the level they say (usually a CEFR level) and merely differ as to which point of the level they are at.

**Remember**

Each level (in particular each CEFR level) represents a range of ability and knowledge.

▶ Chapter 5

If the results of practice exams, whether the first or the second, are disappointing, make sure that you are familiar with the specifications and are giving your learners instructions as they are given in the exam. The more you and your learners know about the exam, the more valid and reliable their results will be. It is unlikely that learners will read specifications, but if you know exactly what is being tested, you will be able to help them as much as possible. This will apply particularly to open questions and tests of productive skills. Are you teaching what the exam is testing? Are your learners being given the opportunity to practise the right things? Make sure that they are familiar with the instructions so that they do not have to spend a lot of time reading them and understanding them during the exam. Give them information and practice in time management so that they use all the time effectively, thinking about as well as checking their answers. As well as being more reliable, the results of tests will also be better the more test-takers know about the test.

## 3 Test tasks and classroom tasks

Test tasks and classroom tasks are designed differently and have different aims. A course preparing learners for an exam should not just concentrate on practice exams or other practice material. In general exam material is not suitable for teaching and will soon become boring as it is not generally topical, controversial or even particularly interesting.

▶ Chapter 7

Exam material for achievement and proficiency tests is chosen to test the product of learning and not the process. Teaching is concerned with the process. However, having said this, there are ways in which sample material from exams, if you have enough of it, can be used in class. It will most probably need to be adapted slightly for this, but can for instance be used in the following ways.

Reading texts can be used with classroom activities, such as conversational warm-ups on the topic in general, detailed discussion of or debate on the topic, information gap

exercises, quizzes etc. You can add to the text with additional information on the same topic, either from the press, the Internet or course books. Choose topics which are likely to come up in the exam. You should be able to find information on the topics which may be used in the exam in the specifications. The same applies to the topics you use for listening activities.

As well as choosing the topics, for reading and listening tests, using the types of texts in class will help to familiarise learners with these for the exam. In this case, you will have to define or identify the types of texts and find as many examples of these as you can. If you have to find and produce material and activities yourself, it is worth building up an exam preparation collection which you can use again and again. You can of course include topical subjects for classroom use or topical aspects of general subjects, but point out to learners that it is unlikely that highly topical issues will be addressed in the exam.

It is easiest to prepare for tests if they include tasks similar to real-life or course tasks. If the listening comprehension part of the exam uses recorded material with a variety of accents, your learners should have been exposed to these in class. You will need to start this early on in the course so that they get enough practice in it before the exam.

This particularly applies to tests of productive skills. If, for instance, in the oral exam, candidates are examined in pairs or groups, it will be essential that you do pair or group work in class. Similarly, if you do mainly group work in class and the exam consists of an interview with an examiner, you may have to adjust your teaching to this. For tests of speaking and writing, familiarise learners with the criteria and encourage them to evaluate their own and their fellow-learners' performances. This can be done with writing and with speaking. You cannot expect them to become examiners, but getting them to think about their own and others' strengths and weaknesses will help them in the exam.

Go through the exam step by step. It is not necessary to do this all at once as this could become tedious, but concentrating on a different part in each session and dealing with it in detail will help learners. Make sure that they ask you all the questions they have before the exam itself and find out the answers if you do not know them already. Ask learners to make a collection of tips which they think will be useful in the exam and add some of your own. Some exam boards publish lists of tips for candidates which tell them how to do well in the individual parts of the exam.

## 4 Skills (business and language) being tested

**Remember**

The judges' decision is usually final in external exams. Unless a mistake in calculation has been made by the exam board, there is little chance that an appeal will be successful.

Make sure that you advise your learners properly. It is important not to discourage them, but also not to raise their hopes unrealistically. Remember that your judgement of how good their BE is may not be the same as the judgement of the examiners or raters. Whatever you may think about your learners' ability will play no part in their final result. You may feel that your view is more comprehensive and more realistic, but if the test-taker does not fulfil the exam's or the examiners' expectations, then he or she will not get a good result, whether one is deserved or not and whatever you think.

▶ Chapter 6

There is plenty of published teaching material (not only exam practice) available to prepare for (and not only practise) all the main exams offered. As well as this, there are books with ideas on how to teach exam classes. Some exam boards publish tips for candidates, and some also provide "official" teaching material. There is additional

information on specifications, as well as tests and tasks for specific skills in other chapters of this book.

▶ Chapters 4, 6 and 7–12

Your learners should not expect you to spoon-feed them with all the information they need. Point out to them where they can find information on the exam and let them decide how much and what they want to practice. Help them to find material to practice with, podcasts, texts and exercises at the right level and of the right kind. You can use their feedback on the usefulness of the material you give them for other courses and learners.

## Your classroom tests

If you are setting tests yourself, you will find things much easier than if you are preparing your learners for an exam over the contents and marking of which you have no control. However remember that in general learners will not want to be patronised and do a test which does not give them a true reflection of their abilities. Progress tests during a course can serve a useful function if they are set well and used properly. Achievement tests should be set to reflect the highest level of achievement you can expect or the level of the CEFR you are aiming at. Although the preparation for a classroom or end-of-course test will not be as intensive as for an external exam, nonetheless some of the same principles apply. There are also additional things you can bear in mind.

Learners should know what the consequences of the results of a test will be, whether these are serious or not. A test must have some point in order to be valid, otherwise there is no reason for doing one. If you or your learners want to have regular progress tests during a course, there is no reason why you should not use the same format each time. That way learners can concentrate on the content rather than on the tasks. A test does not have to be creative and fun in the same way as classroom activities do. Do not however be tempted to merely copy exercises out of the course material you have used. If learners have done the tasks in the same form before, they will feel insulted. The test should be a representative sample of all that has been done during the period of time covered. Testing is not the same as learning or practising.

To establish what you need to test, involve learners in the testing process. Find out from them what they need to know and what they think they do know or can do well. Ask them what they would like to be tested on. There is no reason why their wishes should be less valid than your ideas.

If you are testing them on productive skills, discuss criteria (before setting them finally). This is not a question of letting learners decide on their results, but of making sure they understand why you are testing in the way you have chosen. Explaining and justifying your criteria will also help you to make sure that they are the right ones for the tasks. Make sure they (and you) understand why you have chosen these criteria.

▶ Chapter 13

How much preparation you do for a test will depend a great deal on you and your learners, but some learners are also helped by examples of performance and reasons for marks. If you do classroom tests or practice tests, post-test feedback is particularly important so that learners do not make the same mistakes again. This not only applies to the content of the test, i. e. the language, but also to strategic mistakes or mistakes of time-keeping. As tests are usually set for groups, this will mean that only a certain amount of feedback can be general and much will have to be individual. You will have to set aside time for this or deal with it by email.

The most important thing to remember is that feedback must be realistic, useful and motivating. Giving learners messages and information they can use may mean that they decide not to take the test or to take it at a later date. This can be motivating if they feel that their chances of passing later are higher and they know how to achieve the best possible result.

## 5 Dealing with learners' worries

Learners will probably be more nervous about taking an external exam than about taking a classroom test, especially if the results are important to them and if they have paid money to take the exam. This is quite natural and should not be ignored. Many learners will be influenced by their earlier experiences of tests and exams (at school) and assume that the experience will be unpleasant. Remember the more information you have, the better you will be able to help them to benefit from the experience of taking the exam and get the best result possible.

### Telling them how to get full points

It can be assumed that tests-takers want to get the best result and the highest grade possible. You clearly will not be able to ensure that they answer all the questions correctly, but you can give them useful tips which will help them to avoid any non-language mistakes. If they are nervous, these are things which many test-takers do not do, especially if they have been out of a learning or test-taking environment for some time. The tips range from very general ones, such as reading the questions carefully and planning the time available best, to more specific ones, deciding in what order to answer the questions and planning and/or writing an outline for a written task, to very specific ones, such as trying to answer multiple choice questions without reading the options available, and similarly, filling in gaps. Finally, keeping to the instructions and the tasks and checking answers at the end is definitely important, but something test-takers often do not do.

■ Many of the methods, strategies and tips useful for preparing learners for tests and exams are not the sort of thing which teachers are used to doing in class. It is important to know why you are doing these things and strike a balance in lessons between teaching and test preparation. Learners will not be helped if you ignore the fact that they will be taking a test as part of a course but nor will it be useful if all they do is prepare for the test. If the test fulfils the criteria for a good test, teaching learners in such a way that they are well-prepared for it and get the best results will be easy and satisfying for all concerned. ■

**To think about:**

- How often do you do tests in class?
- How much do your learners know about the test(s) they have to do?
- How autonomous are your learners as far as test preparation is concerned?

CD-ROM

# Concluding remarks

If you have got this far, several things should have become clear.

Teaching and testing are closely related but they are not the same thing. For instance, most tasks which are good for teaching purposes cannot be used in the same form in tests. They usually need adjustment to make them valid and reliable and they also generally require additions, for example in the form of instructions for the test-takers and marking criteria for the examiners or raters.

Testing is not theoretical. In fact it is quite the opposite, although, as with most complex fields, the theory is extremely useful. All good tests are based on rules and principles which should be made available to everyone involved in the testing process. It is important to bear basic principles of testing in mind, but it is the transfer of this theory into practice, however, which is the most interesting and important part of testing.

Setting tests is not difficult. It may be complex as well as a challenge for many BE trainers, but a great number of the qualities of good teaching can be applied to test setting. For instance, there are many similarities between setting tests and designing courses. A systematic approach and an attitude which takes the interests of all the stake-holders into account are the two best things to have when setting tests. When you start setting a test or choosing a test for your learners, you should also be goal-oriented and structured.

Testing takes time. There is far more to setting or choosing tests and testing than just thinking about the questions and answers. You have to develop a system and apply this. Depending on the type of test, different things will take up more or less time, but there will be no way round spending time on testing if you want to take it seriously and produce good tests.

Test-setting and testing should not be done alone. Any test involves more groups of people than just the setter, and it is these groups who should be involved in the process of test design, setting and administration. Some groups who should be involved are test editors, learners for pre-testing and proof-readers.

Every test makes an impact and has consequences, even if these may not be immediately recognisable. Before you think about giving a test, whether a test you have set yourself or a published test, think about all the consequences the test may have, on the learners, on you, on your teaching, on the institution and on all the stake-holders involved. You should always have a good reason for doing a test and be aware of its effects.

Setting a test is not the same as acting as an examiner or rater. Some of the qualities of good teachers, such as creativity, a desire to help or a willingness to ignore mistakes may not be helpful for an examiner or rater. Examiners need to be able to stick to rules and standards in order to be able to act as fairly as possible. Teachers may be able to bend rules now and again but this should not happen in tests.

Taking a test may not be enjoyable for test-takers but neither should it be a tortuous experience which they look back on with displeasure or regret. It should always have some beneficial impact on all those involved. Neither should setting a test be something which BE trainers regard with dread. There are many benefits to everyone to be gained from high-quality tests and this is the only sort of test which professional BE trainers should produce.

**Happy testing!**

# GLOSSARY

The glossary contains terms which may be unfamiliar to general BE trainers or are used with a specific meaning in a testing context. Not all the terms in the glossary have been used in the book. Explanations of terms in italics can be found in the glossary.

## A

**achievement test** – a *test* of what learners have been taught/learned. This can either be of specific language items or certain language objectives.

**administer** – the work involved in organising and running a test. This may include invigilation and issuing results and does not require any particular qualifications.

**analytic marking scheme** – a *marking* scheme which breaks down a total *score* into different areas.

**anchor item** – a tested item which is included in two or more *test*s and produces known results for the group of *test-takers*. Anchor items sometimes form part of a new version of a *test* and provide useful information about *items* and *test-takers*, particularly the level.

**answer** – what *test-takers* write in a *test*.

**assessment** – the most general word for any form of judgement of performance. This includes *tests* and *exams*.

**authenticity** – based on a realistic and acceptable view of language. This is one of the main criteria for a good *test*.

## B

**band** – level in one *criterion* or skill.

**benchmarking** – establishing a standard against which to measure performance or comparing performance against an external standard.

**bias** – occurs when a *test* favours one group of *test-takers* over another, possibly for reasons of culture, age, sex, race, etc.

## C

**calibration** – the process of establishing that all *items* in a *test* are at the required level of difficulty or that all *examiners* are using standards in the same way to *mark*.

**candidate** – the person taking a *test*, usually in a *high-stakes exam*. The more general term *test-taker* is used in this book.

**cheating** – breaking the rules of a *test* to get an unfair advantage.

**clerical marking** – *marking* done by people without any specialist knowledge or qualifications using a set system and list of correct *answers*.

**closed question (fixed response format)** – questions to which there are only a small limited number (often only one) of correct *answers*.

**cloze test** – consists of a written *text* with words deleted, either randomly or according to a system. The *test-taker* has to provide the words.

**computer adaptive test** – a *test* which automatically adjusts the questions to individual *test-takers'* responses and provides further *questions* at an appropriate level of difficulty.

**computerised marking** – automatic *marking* usually using an optical mark reader (OMR) or a scanner.

**consequential validity (impact)** – the effect the existence of a *test* or its results have on individuals or groups.

**construct validity** – how far a *test* gives information on a previously defined view of language.

**content validity** – how far a *test* covers the content it should. This may include information, grammar or vocabulary.

**correlation** – describes the relationship, usually statistically, between two different things, e. g. the level of a *test-taker* and the *answer* to a particular *question*.

**criterion** – the aspect of a language skill to be *marked* as well as the definition for a particular score.

**criterion validity** – how far a *test* measures the same thing as another *test* or *assessment* procedure or corresponds to trainers' views of *test-takers'* ability.

**criterion-referenced test** – describes the *test-takers'* results against a set of *criteria*, regardless of the other *test-takers* and their results.

**c-test** – developed in Germany, similar to a *cloze test* but following particular rules for the removal of parts of words.

**cut point (cutscore)** – the score at which a *test-taker* reaches a particular level or passes or fails.

## D

**descriptor** – the description of a *band* or levels in a particular skill.

**diagnostic test** – a *test* to establish learners' strengths and weaknesses.

**discrete item testing** – using individual *items* rather than groups of *items* in a context.

**discrimination** – how far an *item* distinguishes between weaker and stronger *test-takers*. This can be calculated manually or using a computer programme.

**distractor** – an *option* in a *multiple-choice question* which is not the right *answer* (*key*).

**domain** – area of language use, e. g. occupational.

## E

**exam(ination)** – a formal *test* often recognised nationally or internationally and with *high-stakes* involved.

**examiner** – the person who *marks* a *test*, usually of productive skills. The term is used in this book for the person who *marks tests* of speaking.

**expert judgement** – used in the *test*-setting process to evaluate a *test*.

## F

**face validity** – the appearance of a *test* and whether it looks as if it is testing what it is supposed to test.

**facility** – description (usually statistical) of how easy or difficult a *test item* is. It can be simply the proportion of *test-takers* who answered it correctly.

**formative evaluation** – *assessment* or testing carried out to evaluate the process of teaching and learning in order to improve it.

**G**

**grade** – the result of a *test* expressed as a relation to the whole, often a percentage or a description.

**H**

**high-stakes** – used to describe a *test* where the results have an important influence on the *test-taker's* life.

**I**

**impact** – the effect the *test* has on teaching and learning, its interest and importance to *test-takers* and other *stakeholders* and how far the results are recognised.

**integrative testing** – testing of more than one skill at the same time.

**interactiveness (interactivity)** – how far *test-takers* are able to understand what they have to do in the *test* easily.

**intra-rater reliability** – how far different *raters mark* in the same way.

**item** – each part of a *test* which is given a separate mark, e. g. one gap in a *cloze test*, one *multiple-choice question*, one sentence to be transformed, one *question*.

**item analysis** – analysis of the features of an *item* or *question* as to its level of difficulty, *correlation* with other parts of the *test* or *test-takers'* features etc.

**item bank** – a collection of *items*, *questions*, *tests* and *tasks* which can be used again in *tests* in different combinations or with modification.

**K**

**key** – the right answer in a *multiple-choice question*. Also used to mean the list of right answers in a complete *test*.

**L**

**level descriptor** – the description of a level (e. g. in the CEFR) as well as the description for a *band* in a *test*.

**low-stakes** – used to describe a *test* which has little effect or *impact* outside the particular environment.

**M**

**marking** – assigning a score to a *test-taker's* responses in a *test*. It is sometimes (wrongly) referred to as "correcting".

**multi-matching** – a type of *test* where all the answers are provided (sometimes more than the number needed) and the *test-taker* has to allocate them to the appropriate place or gap.

**multiple-choice question** – a *test item* with several choices (usually three or four) where one (or sometimes more) is right.

**N**

**norm-referenced test** – a *test* where *test*-takers are graded according to the results of the group rather than external *criteria* or standards.

**O**

**objectivity** – often considered one of the features of a good *test*. An objective *test* measures *test-takers'* performance or knowledge independently of *bias*, feelings or beliefs. Objectivity is obtained through standardisation of *tasks*, *items* and procedures.

**option** – one of the choices in a *multiple-choice* question, including the right answer (*key*) and the wrong ones (*distractors*).

**P**

**performance test** – a *test* where the *test-taker* has to actually do something to demonstrate ability, rather than simply produce evidence of knowledge.

**placement test** – a *test* used to put the learners into groups or courses according to their ability and knowledge.

**practica(bi)lity** – how far the setting, running and *marking* of the *test* can take place without unnecessary administrative effort and resources.

**pre-testing** – a stage in the development of *tests* in which *items* are tried out to see how they work in practice.

**proficiency test** – a *test* to ascertain whether a *test-taker* is able to perform in certain circumstances. It is not related to a specific teaching programme but often used as a predictor of future ability.

**Q**

**question** – an *item* in a *test* (whether followed by a question mark or not).

**R**

**rater** – often used interchangeably with *examiner*. Here it is used for the person who *marks* a *test* of writing. Also known as scorer, *examiner* or even (wrongly!) "corrector".

**raw (score)** – a result in the form of numbers, before it is changed into a percentage or a *grade*.

**reliability** – how far a *test* is free from errors of measurement or errors resulting from human weakness and consistent over different versions.

**reporting** – how far the *test* setting process and procedure as well as all other aspects of the *test* are made available to those who are interested.

**response** – answer given by a *test-taker*, usually in a *test* of speaking or writing.

**rubric (instructions)** – the instructions given to *test-takers* in the *test*. They can include information on the order in which questions are to be answered as well as where *answers* are to be *marked*. The term is sometimes also used to mean *criterion*.

**S**

**set** – the process of designing and writing a *test* or parts of a test. It may involve choosing topics, writing texts and constructing *questions*. Test setting requires linguistic qualifications and experience.

**specifications** – information about a *test*, including what is tested, how it is tested and details of *test format* and *marking*.

**stakeholder** – anyone who has an interest in the *test* and its results, including *test-takers* themselves, prospective employer companies, institutions, parents etc.

**stem** – the part of a *multiple-choice item* which provides the setting for the *options*.

**sub-test** – section or part of a *test*, often testing one skill or area.

**summative evaluation** – carried out in the final stage of a process to get feedback on the process, course or teaching.

**syllabus (curriculum)** – the basis of the teaching or the *test*, the content of what is being taught and tested.

## T

**task** – part of a *test* usually consisting of several *questions* or *items* in one context or a more complex performance by the *test-taker*.

**test** – the blanket term in academic descriptions for all forms of *assessment* and evaluation. When used in comparison to the term *examination* it is considered less formal or serious.

**test construct** – what is being tested. This can be the view of language, the skills involved or the knowledge.

**test content** – what is in the *test* in terms of information and aspects of language.

**test format** – the structure of the *test*, the number of tasks, questions, timing etc.

**test-taker** – the person taking the *test*. The term "candidate" is also commonly used.

**text** – used for reading and listening, the input which forms the basis of a *test* and on which *questions* are set.

## V

**validity** – how far the *test* actually tests what it sets out to do. The most important feature of a good *test*.

## W

**washback (backwash)** – the effect a *test* has on the learning and teaching process (closely related to *impact*).

**weighting** – the relation of different parts of a *test* to each other in importance and score.

# Further information

## Books

This list includes most of the books quoted or referred to in this publication and a few more general books on testing. No extremely detailed or theoretical books have been listed here as the aim is to provide reading which gives a general introduction to the main issues involved.


Allen, J. P. B./Davies, Alan (eds.) (1977): The Edinburgh Course in Applied Linguistics Volume 4: Testing and Experimental Methods. OUP, Oxford

Bachman, Lyle F./Palmer, Adrian S. (1996): Language Testing in Practice. OUP, Oxford
Bachmann, Lyle F. (2004): Statistical Analyses for Language Assessment. CUP, Cambridge
Baxter, Andy (1997): Evaluating Your Students. Richmond, Slough
Brieger, Nick (1997): Teaching Business English Handbook. York Associates, York

Carroll, Brendan J./Hall, Patrick J. (1985): Make Your Own Language Tests. A Practical Guide to Writing Language Performance Tests. Pergamon Institute of English, Oxford
Carroll, John B. (1961): Fundamental Considerations in Testing for English Language Proficiency of Foreign Students. Washington Center for Applied Linguistics, Washington
Cheng, L. & Watanabe, Y. (2004): Washback in Language Testing. Lawrence Erlbaum Associates, Mahwah
Council of Europe (2001): Common European Framework of Reference for Languages: learning, teaching, assessment. CUP, Cambridge

Daller, Hellmut/Milton, James/Treffers-Daller, Jeanine (2007): Modelling and Assessing Vocabulary Knowledge. CUP, Cambridge
Davidson, F./Lynch, B. (2002): Testcraft. Yale University Press, New Haven
Donna, Sylvia (2000): Teach Business English. CUP, Cambridge

Frendo, Evan (2009): How to Teach Business English. Pearson Longman, Harlow
Fulcher, G./Davidson, F. (2007): Language Testing and Assessment. Routledge, London
Fulcher, Glenn (2010): Practical Language Testing. Hodder Education, London

Heaton, J. B. (1990): Classroom Testing. Longman, Harlow
Hewings, Martin/Nickerson, Catherine (1999): Business English: Research into Practice. Longman, Harlow
Hughes, Arthur (2003): Testing for Language Teachers. CUP, Cambridge

Kelly, T.L. (1927): Interpretation of Educational Measurements. New World Book Company, New York
Krashen, Stephen D. (1981): Second Language Acquisition and Second Language Learning. Pergamon Press, Oxford

Lado, Robert (1961): Language Testing. Longman, Harlow
Lewis, Michael (1993): The Lexical Approach. LTP, Hove
Luoma, Sari (2004): Assessing Speaking. CUP, Cambridge

McNamara, Tim (2000): Language Testing. OUP, Oxford
May, Peter (1996): Exam Classes. OUP, Oxford
Morrow, Keith (ed.) (2004): Insights from The Common European Framework. OUP, Oxford

O'Malley, J. Michael/Valdez Pierce, Lorraine (1996): Authentic Assessment for English Language Learners. Longman, Harlow

Pit Corder, S. (1973): Introducing Applied Linguistics. Penguin Books, Harlow
Purpura, James (2004): Assessing Grammar. CUP, Cambridge

Underhill, Nic (1987): Testing Spoken Language. CUP, Cambridge
Urkun, Zeynep/Mader, Judith (2009): Recent Approaches to Testing and Assessing Speaking. IATEFL TEASIG (CD-ROM)
Urkun, Zeynep (ed.) (2007): The Common European Framework of Reference for Languages: Benefits and Limitations. IATEFL TEASIG (CD-ROM)

Weir, Cyril (1990): Communicative Language Testing. Prentice Hall, Upper Saddle River
Weir, Cyril (2005): Language Testing and Validation. Palgrave, Basingstoke
Weir, Cyril (1994): Understanding and Developing Language Tests. Prentice Hall, Upper Saddle River

## Websites and downloads

All the information given here was correct at the time of going to press.

**www.alte.org**
ALTE is an association of providers of foreign language examinations, founded in 1990, which includes many of the world's leading language assessment bodies. The members represent the testing of 27 languages and provide standard examinations and computer based language tests. ALTE has also produced various publications, leaflets and newsletters about its work as well as resources for language testers.

**www.alte.org/downloads/index.php?docid=89**
Materials For the Guidance of Test Item Writers.

**www.coe.int/t/dg4/linguistic/manuel1_en.asp**
The Council of Europe website contains the full text of the CEFR and the *Manual for relating Language Examinations to the Common European Framework of Reference for Languages* as well as several related publications.

**http://languagetesting.info**
This site, maintained by testing expert Glenn Fulcher, acts as a reference guide to resources related to language testing on the Internet by providing a history of, and links to, information on language testing, including publications that are freely available on other domains. It also has resources such as articles, features, videos and audio, which are freely available for language teachers, language testers, and students of language testing, applied linguistics, and languages.

**www.ealta.eu.org**
EALTA is a professional association for language testers in Europe with interests independent of those of any other organization. EALTA was set up with financial support from the European Community with the purpose of promoting the understanding of theoretical principles of language testing and assessment, and the improvement and sharing of testing and assessment practices throughout Europe. The website contains information on the organisation as well as some resources. Individuals can join EALTA free and have access to the email discussion list.

**www.iltaonline.com**
ILTA is an international group of language testing and assessment scholars and practitioners. ILTA's purpose is to promote the improvement of language testing throughout the world as well as ethical standards in language testing through its Code of Ethics. The website has links to conferences, journals and resources on testing. ILTA also provides an extensive testing bibliography to members. The discussion list LTest-L is open to anyone.

**http://tea.iatefl.org/**
This is the website of the IATEFL TEASIG (the Testing, Evaluation and Assessment Special Interest Group) and contains information on events for IATEFL and TEASIG members. TEASIG also runs an open discussion list for teachers on testing issues.

## Email discussion lists

These lists are easy to join and hold fairly regular discussions on many different aspects of testing, some theoretical and some practical.

EALTA: www.ealta.eu.org
ILTA: www.lists.psu.edu/archives/ltest-l.html
TEASIG: http://groups.yahoo.com/group/TEASIG/

## Journals

These are just some of the many journals on testing related issues.
Language Testing Journal

http://ltj.sagepub.com/
Practical assessment research and evaluation
www.pareonline.net
English Language Teaching Journal
www.eltj.oxfordjournals.org

# Acknowledgements


## Thanks and dedication

Thanks go to Eric Baber and Stephanie Hempel for their excellent editing support and Isobel Williams for her encouragement and help throughout.

The content owes a lot to my former colleagues at WBT-Weiterbildungstestsysteme GmbH in Frankfurt, Germany, in particular Heinrich Rübeling, Beate Zeidler and Peter Kiefer, who provided me with many insights into test development. Without the discussions we had and the experience I gained from working at the WBT, this book could never have been written.

Very special thanks go to Rudi Camerer, for his invaluable encouragement, support and help then and now.

This book is dedicated to him and my children, Gabriel, Naomi and Jessica.

## Sources

The authors and publisher are grateful to those who have given permission to reproduce the following illustrations, photographs and texts.

**Illustrations:**
p. 19 and 29: Andreas Terglane
p. 27: Cartoonstock/Theresa McCracken
p. 35: Cartoonstock/Joel Mishon
p. 48: Cartoonstock/Jason Love
p. 99: Cartoonstock/Chris Wildt
p. 122: Cartoonstock/Roy Delgado
p. 130: Cartoonstock/Dan Collins
p. 141: Cartoonstock/Aaron Bacall
p. 150: Cartoonstock/Ron Morgan
p. 159: Cartoonstock/Jonny Hawkins
p. 165: Cartoonstock/Andrew Toos

**Photographs:**
p. 55: Paul Thompson Images/Alamy
p. 68: mediablitzimages (uk) Limited/Alamy
p. 79: © Kathy deWitt/Alamy
p. 86: Asia Images Group Pte Ltd/Alamy
p. 107: Paul Cotney/Alamy

**Texts:**
Slides on page 29: Original concept from McNamara, Tim (2000): Language Testing. OUP, Oxford. Slides reproduced by kind permission of elc-European Language Competence
CEFR-Scales on the CD-ROM: Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). Published by the Council of Europe. www.coe.int/t/dg4/linguistic/cadre_en.asp
The author first heard the expression “the end of the line” applied to testing from Maria Sara Rodriguez in Uruguay. It is used in an article in the IATEFL TEASIG Newsletter Spring 2011.

# List of contents/ CD-ROM

# Testing and Assessment in Business English

**Testing and Assessment in Business English** is a handbook and resource pack for Business English trainers and directors of studies who want an easy-to-use overview of the theoretical and practical issues involved in language testing. This book guides the reader through the main areas of testing relevant to Business English professionals. These include designing specifications, choosing a test or an exam, formulating marking criteria and writing test instructions. There are also chapters on testing each of the four skills, as well as on testing vocabulary and grammar.

**Testing and Assessment in Business English** is specifically geared to the daily requirements of busy trainers. The chapters can be referred to separately according to specific needs and interests, or the whole book can be worked through systematically.

**Testing and Assessment in Business English** includes:

- Chapters on the features of good tests, test specifications, types of tests, choosing a test, testing different skills, marking criteria, as well as guidelines on preparing learners for tests and exams
- Further reading tips
- Link lists
- Glossary of testing terms
- CD-ROM with checklists for each chapter, test templates and workshop material

Judith Mader is Head of Languages at the Frankfurt School of Finance and Management, a private business school in Frankfurt, Germany, and Senior Assessment Consultant for elc-European Language Competence, Frankfurt, a language consulting company for companies, publishers, ministries and chambers of commerce internationally.

Cornelsen

ISBN 978-3-06-520080-6

9 783065 200806