LANGUAGE TESTING

A Critical Survey and Practical Guide

David Baker

LANGUAGE TESTING

A Critical Survey and Practical Guide

David Baker

Testing Co-ordinator,
University of Bahrain English Language Unit

Edward Arnold
A division of Hodder & Stoughton
LONDON NEW YORK MELBOURNE AUCKLAND

First published in Great Britain 1989

British Library Cataloguing in Publication Data

Baker, David
Language testing : a critical survey
and practical guide
1. Educational institutions. Students.
Modern language skills. Assessment.
Tests.
I. Title
418'.007

ISBN 0–7131–6538–3

Photoset in Linotron Ehrhardt 10/11½pt. with Univers by
Northern Phototypesetting Company, Bolton
Printed and bound in Great Britain
for Edward Arnold, the educational academic and medical publishing
division of Hodder and Stoughton Limited, 41 Bedford Square, London
WC1B 3DQ by Richard Clay, Bungay, Suffolk

CONTENTS

1 Making tests and making decisions. 1
1.1. Shifting perspectives 1
1.2 Making judgements and using jargon 2
1.3 Testing decisions and procedures 3
A test is a way of arriving at a meaningful decision 3
A test is a substitute for a more complete procedure 4
2 Four models 7
2.1 Language as action vs language as system 7
2.2 Performance-referenced testing 11
Direct testing 11
Indirect performance-referenced tests 16
2.3 System-referenced tests 23
Direct system-referenced tests 24
Indirect system-referenced tests 24
2.4 Resumé 27
2.5 Application of the models 28
3 The psychometric legacy 29
3.1 The origins of modern language testing 29
Psychometric testing 29
Structural linquistics 30
3.2 Details of the psychometric approach 31
The dimensions of proficiency 32
Elements vs situations 33
Discrete-point 'objective' test formats 34
Summary 35
3.3 Evaluating the approach 36
The model 36
The instruments 37
3.4 The pass-mark problem: is norm-referencing wicked? 39
3.5 Summary 40

4 The use and interpretation of statistics 41
4.1 The surrogate judgement 41
4.2 Questions with statistical answers 41
4.3 Describing the population 43
Quantifying the distribution: means and variance 44
Interpreting the statistics 45
Using distribution statistics to make decisions about tests 46
Using distribution statistics – three case studies 46
4.4 Ranking statistics 49
The rank score 49
Percentile score 49
4.5 Item analysis 51
The facility index 51
The discrimination index 52
Distractor analysis 53
Why do item analysis? 53
Measures, scales and lists 54
4.6 The relation between measures – the correlation coefficient 56
The scattergram 56
The correlation matrix 57
The use and interpretation of correlation coefficients 57
Handle with care 58
4.7 Calculating reliability 60
Inter-rater reliability 60
'Hypothetical' reliability 61
4.8 Factorial analysis 61
Looking for patterns in the data 62
Principal factor analysis 63
Principal component analysis 63
4.9 Conclusion – recognizing the limits 63
5 The integrative interlude 65
5.1 The reaction against psychometric testing 65
5.2 The Unitary Competence Hypothesis 66
The evidence from factor-analytic studies 68
The nature of proficiency – the 'pragmatic expectancy grammar' 68
The end of the UCH 70
5.3 The structure of language proficiency – after the UCH 71
5.4 The use of cloze and dictation tests 72
Using cloze tests in practice 73
5.5 Using dictation 74
6 Performance-referenced and task-based test designs 76
6.1 Economy and comparability – drawbacks of performance-referenced testing 79
6.2 The development of direct performance-referenced tests 80
Sampling the criterion performance – choosing the tasks 81
Measuring the direct test performance 84

Making the judgement 88
6.3 Indirect performance-referenced tests 91
Analysis of the criterion performance 92
Sampling the criterion proficiency 94
Test construction 94
Measurement and judgement 95
Indirect tests – advantages and drawbacks 96
6.4 Performance-referenced testing and statistical techniques 97
Direct tests 98
Indirect tests 98
7 Test type and test purpose 100
7.1 Identifying decisions 100
7.2 Testing within language instruction programmes 101
Entrance tests 101
Progress tests 103
Exit tests 104
7.3 Public language examinations 105
Task-based public examinations 106
7.4 Deciding to test – seven key questions 106
References 108
Index 111

1

Making tests and making decisions

1.1 Shifting perspectives

The past ten years or so have seen a number of changes in the practice of language teaching. Some of these changes have been superficial, not having much effect on what teachers do in classrooms. Others have been short-lived fashions. Nevertheless it is possible to see that there has been a change in emphasis. The language teacher used to be in the business of helping the learner to master a 'system'. That is, the goals of language instruction were described in purely linguistic terms and the syllabuses which resulted were basically inventories of structural features organized in order of increasing complexity. The language teacher's task was seen as helping the learner to a gradual mastery of these features. The purposes of the language study were given little importance, since it was assumed that the structural features of the language represented an analysis at a sufficient level of generality to be applicable to all learners, from tourists to nuclear engineers. Syllabus design tended to look inward to the constituents of the language system and how they could be most effectively ordered and taught. The details of what the language would subsequently be used for were not thought to be concerns of the language teacher any more than a typing teacher should worry about what kind of text the students will have to type – learning to type begins and ends with the mastery of a well-defined set of motor skills.

Two shifts of interest occurred which changed this viewpoint. The first was the growth of interest in notional-functional syllabuses. This approach challenged the assumption that the selection and ordering of items for a syllabus should be done on purely structural grounds. It was proposed instead that the perlocutionary force of language items and their meaning relationships could be used as a basis for grouping and ordering them for teaching purposes. The effect on teaching and materials was not always as radical as was sometimes claimed: the first chapter of an elementary textbook is now called 'Introducing yourself' rather than 'The verb TO BE', but a quick glance often shows that the same structural repertoire is presented and practised as before (although items like 'Am I a man?' have probably been removed). Nevertheless, interest in 'functions and notions' did result in a shift of emphasis from the language as a hermetically sealed system towards concerns for the social and psychological dimensions of language use.

The second development which had an important effect was the growth of ESP (English for Specific Purposes). Attempts to produce courses 'tailor-made' for specific groups of learners clearly went against the idea of a single common learning process which all learners underwent. The specification of objectives for these courses contained increasing reference to the use of the language to achieve specific tasks in specific situations. The criteria of success or failure for these learners then began to be seen in terms of the performance of these tasks rather than the mastery of a linguistic system *per se.*

These shifts in emphasis in language teaching have inevitably had consequences for language testing. Testing techniques and theories, however, have been rather more resistant to change than theories about methodology and course design. This is principally because modern language testing is based on principles which, like the old 'structural' syllabuses, take as their starting point a description of the language independent of any particular use of it. The development of tests based on these principles is facilitated by a well-tried set of statistical procedures for constructing and evaluating language tests. Changes in approaches to language teaching inevitably resulted in attempts to develop testing techniques appropriate to the new pedagogy. Unfortunately, problems arise when earlier statistical techniques are extended to these texts based on more recent principles. Advocates of such tests have been forced to develop new procedures for developing and evaluating their test instruments. The legitimacy of these new techniques has been called into question.

The result of this has been to make language testing an area of considerable controversy. Such fundamental questions as 'What makes a test a good test?' and 'How should we go about constructing a test?' will receive quite different answers from adherents to different schools. Procedures acceptable to one approach may be anathema to another and so on.

Those involved with language teaching who have to make decisions about using tests can find all of this very confusing. The aim of this book is to put these issues into perspective and to give the user or writer of language tests the necessary conceptual tools to make sound, informed decisions in this field.

1.2 Making judgements and using jargon

There is a fair amount of specialized terminology used in talking about language testing. Often it has the effect of obscuring rather than clarifying the issues involved. We can usually avoid this by speaking plainly and using special terms with care. There is another difficulty with terminology, however, which is less easy to resolve; when a field is in a state of controversy as is the case with modern language testing, it is sometimes difficult to use terminology neutrally. Thus adopting the concepts and terminology of a particular school of testing tends to 'beg the question' when it comes to discussing the value of the procedures of that school or another.

For this reason any discussion which is intended to make sense of the issues must be conducted using terms and concepts which permit even-handed treatment of the claims and approaches of different schools. Before going on to examine in detail these different approaches to language testing, therefore, I am going to map out some common ground and introduce a few conceptual tools which will enable us to talk about each approach from as neutral a position as possible. This will involve asking basic questions about what tests are for and what kind of relationship they have to the 'real' world.

In the next chapter I shall be proposing a couple of models which should make explicit certain principles which operate in language testing. Clarifying these principles will provide a framework within which the approaches which we will examine later on can be located.

First, however, let us take a look at what testing in general is supposed to achieve.

1.3 Testing, decisions and procedures

Language testing is a complicated subject and much of this complication stems from problems of description and measurement which are particularly acute in linguistic and psychological investigation. It can be instructive therefore to look at other kinds of tests which do not share these particular difficulties. Life is full of tests of varying degrees of formality and important principles can often be seen operating more clearly in non-linguistic tests, where issues are simpler. Extending these principles to language testing can help to think clearly about what tests do and what they are for.

We can start by looking at two fundamental principles which provide a starting point for thinking about the goals of any kind of testing.

1.3.1 A test is a way of arriving at a meaningful decision

Testing is invariably associated with the making of decisions. Whenever something or someone is subjected to a test there is a decision to be made. From checking the oil level in a car to testing a baby's bathwater with the elbow, the results of the test will lead to the choice of a course of action. In the first case the motorist must decide whether to put in more oil or not. In the second case the parent must decide whether or not to put in the baby.

Langage tests also lead to decisions: a placement test, for instance, allows a school to decide in which group a learner will learn most effectively. In the case of language testing, however, this simple truth is obscured by the fact that not all language tests are tests in the real sense of the word. A familiar example is the end-of-year test in the disreputable private language school An end-of-year test should serve to decide whether the learner can pass up to the next 'level'. In certain schools, however, all learners pass to the next level whatever their performance in the first test (the school needs their fees). In this case it is easy to see that this procedure is not really a test at all since the results will change nothing. It is perhaps best regarded as a ceremony, a cathartic ritual to be undergone before the holidays. The person responsible for writing such a test can save himself a lot of the work involved in constructing a real test, since all that is necessary is that the exam be difficult and traumatic and have some vague relationship to the course the learners have followed.

A similar observation can be made about the so-called progress test. In theory a progress test can guide a teacher's decisions about his teaching or the syllabus-designer's evaluation of his programmes. Often, however, its sole purpose is as a goad to encourage regular revision on the part of the learners. Such motivating devices are useful but should not be confused with tests proper. The writer of such 'tests' will be able to write more effective motivating devices once freed from the notion that what is to be written is a test.

The 'decision' criterion can be used to decide whether testing is necessary at all in a given situation. By asking 'what decision do I need to make about these learners?', we can discover whether we need a real test, a ceremony, a goad or nothing at all. Although there is much that could be said about the construction of goads and ceremonies, what follows refers to language tests in the sense outlined above, i.e. procedures that, at least potentially, facilitate decision-making.

If we decide that we need a real test, identifying the decision that needs to be made is an important first step in constructing or choosing an appropriate instrument. If we discover that we do not need a real test, the operation of this criterion may save a lot of time and expense. Appreciation of the close link between testing and decision-making enables the test user or writer to approach the task of evaluating a group of learners with a much clearer idea of what kind of test is needed, if indeed a test is needed at all.

1.3.2 A test is a substitute for a more complete procedure

In the last section we were concerned with what tests are for, what purpose they serve. It was concluded that testing permits the making of decisions. We now have to look at the relationship between the economy of a test and the confidence which can be placed in its results.

Let us go back to the example of testing the oil level of a car with the dipstick. This test is quick and easy, and in general there is no reason to doubt that the level indicated faithfully reflects the volume of oil in the engine. On the other hand, the suspicious motorist always has the option of draining the oil from the engine and measuring it directly. This is much less convenient but, being more direct, eliminates any errors due to faults with the dipstick. There is a trade-off here between ease of administration of the test and the confidence which can be placed in its results. Thus a placement test consisting of an oral interview, writing tasks and various other sub-tests will be less likely to lead to misplacement than a twenty-item multiple-choice test; but it involves a lot more time and trouble.

It is possible to take this idea to an absurd extreme which, however, illustrates an important principle. If a highly sceptical motorist suspects that even draining the oil from the car does not allow him to decide whether to add oil or not (perhaps the volume stipulated in the manual is wrong), the option remains of applying the 'acid test': he can drive the car until the engine starts to complain. At that point he can be 100 per cent certain that it is time to add oil. Similarly the parent who has no faith in the 'elbow test' for the baby's bath water can put the child in and observe the results! In both of these cases, although complete confidence can be placed in the results of the procedures, there is the risk of very undesirable consequences.

It is easy to see that the dipstick and elbow tests serve as substitutes for the more extreme procedures and that we are usually prepared to forgo complete certainty in the results in return for ease of administration. This observation can be generalized to all kinds of testing: a test is always a quicker or easier substitute for a more complete decision-making procedure. This procedure can be called the **criterion procedure**. The criterion procedure is always more difficult or inconvenient than the test procedure but it is the hypothetical performance of the subject during the criterion procedure which the test procedure is designed to reveal.

This is easy to see in other examples drawn from outside language testing. Brick manufacturers, for example, have to decide whether each batch of completed bricks can be sold for building purposes, or whether adjustments need to be made to the manufacturing process. They normally take a sample of bricks from each batch and test them to destruction in a press. This is more convenient than the criterion procedure which would be to build the entire batch into a wall and observe their performance over a period of years. In spite of potential problems with the test (in this case, problems of sampling among others), the manufacturer feels justified in extrapolating from the results of the test to the hypothetical results of the criterion procedure.

Language tests also illustrate this principle. We have already seen that the function of the placement test is to decide which group of learners would be most suitable for a given student. The surefire way of placing a learner in a school (the criterion procedure), would be to put him in a class and see how he gets on, moving him if necessary. This method will eventually guarantee correct placement but is time-consuming and inconvenient. The placement test is a substitute for this criterion procedure. As anyone who has ever used a placement test knows, the results are not always satisfactory but the gain in time and convenience usually makes it preferable to the criterion procedure of letting students 'shop around' the classes. The saving in time and expense is even greater in the case of university entrance exams such as the TOEFL test in the USA or the British Council ELTS tests used by British universities. The function of these tests is to allow universities to decide if the English proficiency of a candidate is adequate for following a course of study. The criterion procedure for deciding this would be to let the candidate start a course and monitor his or her performance. Clearly, considerable time and expense would be wasted in the cases of those candidates who turned out not to be sufficiently proficient. Although the results of the tests may not permit complete confidence in decision-making (maybe the exam excludes students who could, in fact, have coped with their courses, and *vice versa*), the saving of time and money makes the risk worth taking.

Looking back over these examples, from the elbow test, through the dipstick and placement tests to university entrance exams, we can see that each is a short-cut to information about future or hypothetical performances. In each case there is a price to be paid in terms of the confidence with which extrapolations can be made. Clearly in the design of any kind of test a prime consideration must be the minimizing of this price by ensuring that the judgements which are made during the test procedure correspond as closely as possible to those that would be made during the criterion procedure. This involves ensuring that the test and criterion procedures have features in common and that these features can be adequately measured in order to arrive at a judgement. *Which* features of the criterion procedure need to be simulated in the test procedure and how they can be measured is generally much more difficult to specify with respect to language tests than other kinds of testing. In the example of brick-testing, for instance, the feature which both the testing procedure and the criterion procedure have in common is the application of a compression load to the brick. Other features of the criterion procedure (e.g. the covering of the brick with mortar) are not judged to be worth reproducing in the test situation. In constructing a university entrance examination, however, it is not so easy to identify the key features of the criterion procedure: which aspects of a student's language proficiency are crucial to future academic success is not at all clear in the absence of an adequate theoretical description. The adequacy of the test as a ground for decisions may be compromised by failure to specify these features correctly.

The extent to which a test procedure is an adequate basis for decision-making is a question of its **validity**. In the next chapter we will be addressing the problem of validity in its various aspects.

2

Four models

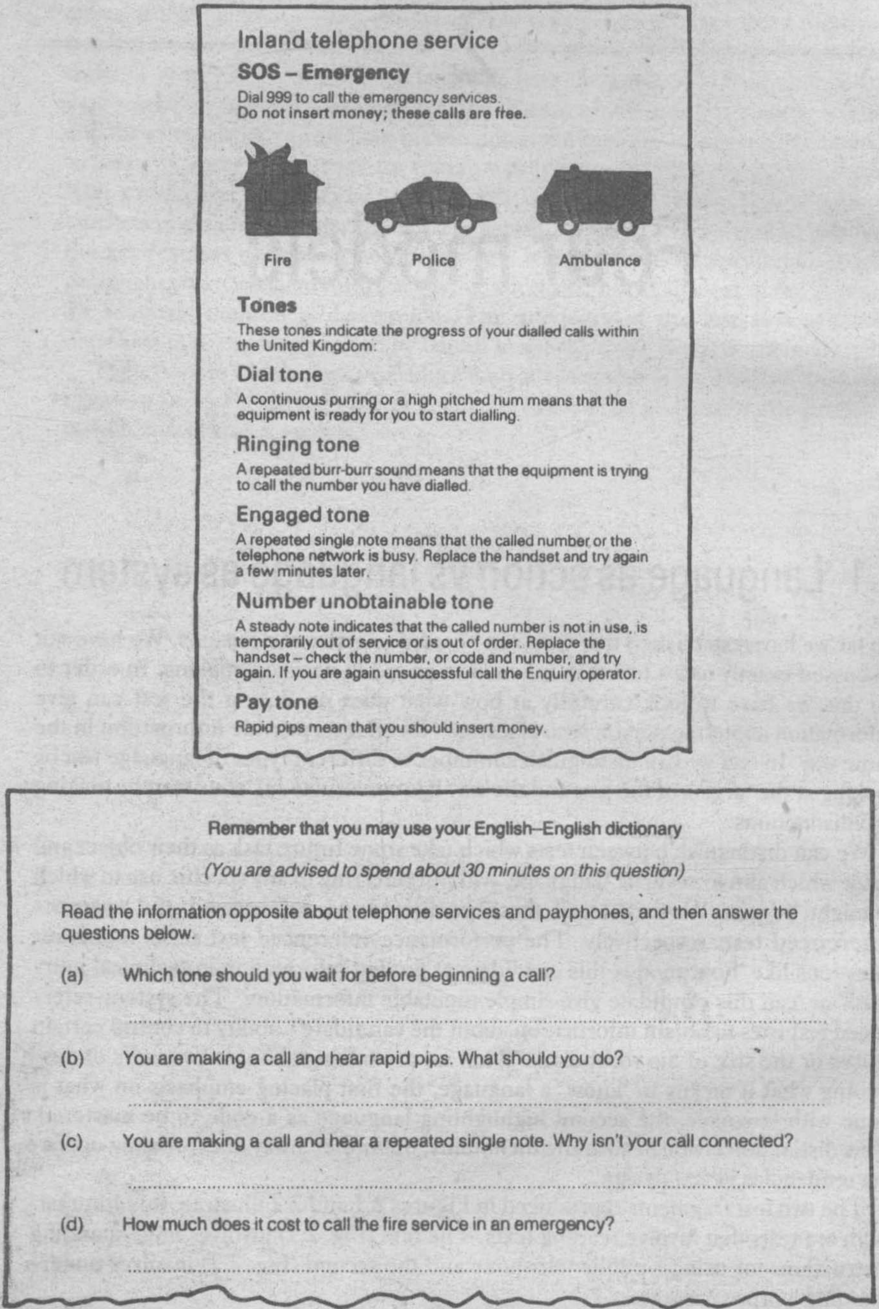
2.1 Language as action vs language as system

So far we have established that tests can be used to arrive at decisions. We have not discussed exactly *how* a test may function as an aid to decision-making. In order to do this we have to look carefully at how what goes on during the test can give information about the person who is tested. Not all tests provide information in the same way. In fact we can distinguish a number of different types of language test by looking at the targets of the test and the way it is constructed. Let us start by making two distinctions:

We can distinguish between tests which take some future task as their object and those which aim to evaluate 'language' without referring to any specific use to which it might be put. We might call these **performance-referenced** and **system-referenced** tests respectively. The performance-referenced test seeks to answer questions like 'how good is this candidate at finding information in technical journals?' or 'can this candidate give simple timetable information?' The system-referenced test tries to obtain information about the candidate's ability to control certain tenses or the size of his vocabulary. What we are talking about is two ways of describing what it means to 'know' a language, the first placing emphasis on what is done with language, the second highlighting language as a code to be mastered. This distinction is not an absolute dichotomy, but rather a way of expressing opposing tendencies in test design.

The two test fragments reproduced in Figures 2.1 and 2.2 illustrate this contrast. Both are tests that involve reading texts. The first (Fig. 2.1) involves understanding instructions for using a public telephone and the second (Fig. 2.2) involves understanding a prose passage.

The first has been designed with a particular performance in mind and would give information about a candidate's ability to perform that specific task. At the same time it would be less justifiable to extrapolate from this test performance to



Inland telephone service

SOS – Emergency

Dial 999 to call the emergency services.
Do not insert money; these calls are free.

Fire
Police
Ambulance

Tones

These tones indicate the progress of your dialled calls within the United Kingdom:

Dial tone

A continuous purring or a high pitched hum means that the equipment is ready for you to start dialling.

Ringing tone

A repeated burr-burr sound means that the equipment is trying to call the number you have dialled.

Engaged tone

A repeated single note means that the called number or the telephone network is busy. Replace the handset and try again a few minutes later.

Number unobtainable tone

A steady note indicates that the called number is not in use, is temporarily out of service or is out of order. Replace the handset – check the number, or code and number, and try again. If you are again unsuccessful call the Enquiry operator.

Pay tone

Rapid pips mean that you should insert money.

Remember that you may use your English–English dictionary

(You are advised to spend about 30 minutes on this question)

Read the information opposite about telephone services and payphones, and then answer the questions below.

(a) Which tone should you wait for before beginning a call?

....

(b) You are making a call and hear rapid pips. What should you do?

....

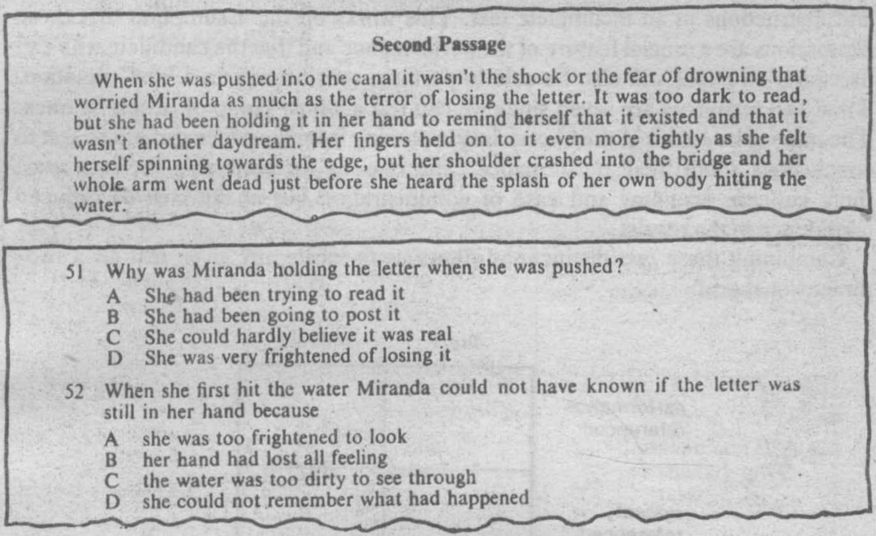
(c) You are making a call and hear a repeated single note. Why isn't your call connected?

....

(d) How much does it cost to call the fire service in an emergency?

....

Fig. 2.1



Second Passage

When she was pushed in:o the canal it wasn't the shock or the fear of drowning that worried Miranda as much as the terror of losing the letter. It was too dark to read, but she had been holding it in her hand to remind herself that it existed and that it wasn't another daydream. Her fingers held on to it even more tightly as she felt herself spinning towards the edge, but her shoulder crashed into the bridge and her whole arm went dead just before she heard the splash of her own body hitting the water.

51 Why was Miranda holding the letter when she was pushed?

A She had been trying to read it
B She had been going to post it
C She could hardly believe it was real
D She was very frightened of losing it

52 When she first hit the water Miranda could not have known if the letter was still in her hand because

A she was too frightened to look
B her hand had lost all feeling
C the water was too dirty to see through
D she could not remember what had happened

Fig 2.2

performance of other kinds of reading tasks. The second example is more general in its applicability but does not give information about any specific type of performance. This tendency to go for increased generality by limiting the domain of a test to linguistic features is typical of early work in language testing (see Chapter 3 for a discussion of this). Performance-referenced language tests, in contrast, are a more recent development. Which kind of test is more useful or appropriate will depend on the nature of the group to be tested. It is up to the user/writer of language tests to decide how generalizable the results of the test need to be and how specifically the potential or future performances can be identified. In general the most confident decisions can be made on the basis of performance-referenced tests but only if the candidates being tested share the same goals and destinations and these can be clearly specified in advance.

Cutting across this distinction is a second distinction between tests whose relationship to their object is **direct**, and those which involve a process of analysis in their construction and are therefore **indirect**.

Using the expressions introduced in the last chapter we can say that in a direct test the test procedure is very similar to the criterion procedure, whilst in an indirect test, features have been abstracted from the criterion procedure.

By way of an example, consider two ways of assessing a candidate's ability to explain how to operate a cassette recorder. The direct way is give him the machine and have him give instructions to an interlocutor. This method has the drawback of being expensive in time resources since only one candidate is tested at a time. On the other hand, if the task is performed satisfactorily, then we can be fairly sure that the candidate will be able to carry out this and related tasks in the future. The alternative method is to have the candidate write the instructions, perhaps filling in key phrases and instructions in an incomplete text. This works on the assumption that these expressions are a crucial feature of the performance and that the candidate who can use them on paper will also be able to perform satisfactorily in a 'real' situation. Time and resources are saved since we can test a whole group of people at once. The price to be paid is in the uncertainty in passing from the paper and pencil test to conclusions about 'real' performance. The reasons for preferring indirect tests, then, concern economy and ease of administration but at the cost of reduced confidence in the results.

Combining these two distinctions allows us to locate any given test on a two-dimensional grid:

	direct	indirect
performance-referenced		
system-referenced		

Fig 2.3

Before going on to look at these test types in detail it is worth sketching out what kind of tests fall into each category.

Performance-referenced language tests owe their development to the desire to have information about what a testee can actually do with his language proficiency. They are of fairly recent origin (although in the fields of vocational and professional training this approach to evaluating ability has a long pedigree and many decisions, from the certification of apprentices to the appointment of civil servants, are taken on the basis of simulation-based tests). Into the **direct** category of such tests come so-called 'communicative' tests in which the test situation is supposed to simulate as closely as possible occasions of authentic language use. The **indirect** tests aim to provide the same information, not by exactly simulating the language performance in the test but rather by breaking it down into more easily testable components. Examples include university entrance tests such as the JMB examination and British Council ELTS test.

System-referenced tests are older in origin. Their aim is to provide information about language proficiency in a general sense without reference to any particular use or situation.

The **direct** system-referenced test is exemplified by the very traditional testing devices of composition and oral interview when these methods are used as ways of getting a sample of language out of the candidate in order to assess its acceptability according to purely linguistic criteria such as grammaticality, vocabulary size, etc.

The **indirect** category includes most public language tests produced since the war: information is required about the testee's general language proficiency (without reference to any particular use or purpose). Rather than evoke directly a sample of language, as in the oral or composition methods, this information is acquired indirectly. Multiple-choice 'grammar' questions and vocabulary quizzes are all examples of this kind of test.

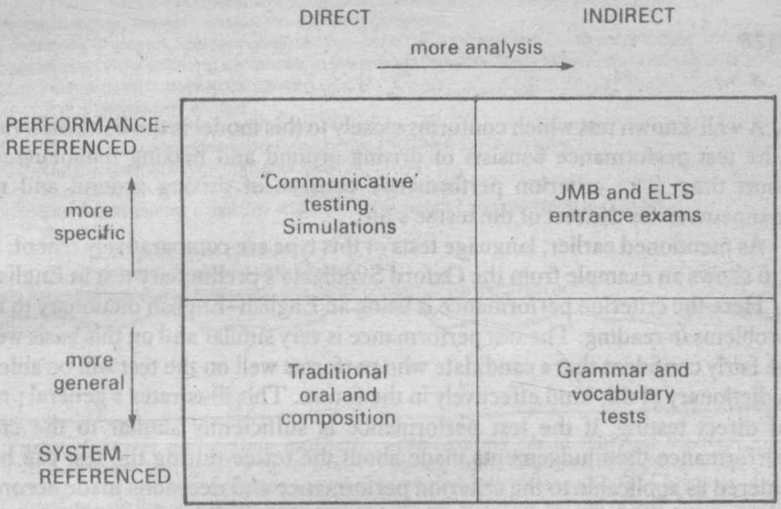

DIRECT
INDIRECT
more analysis
PERFORMANCE-REFERENCED
more specific
more general
SYSTEM-REFERENCED
'Communicative' testing. Simulations
JMB and ELTS entrance exams
Traditional oral and composition
Grammar and vocabulary tests

Fig 2.4

We now have to look in detail at the types of test which have been identified above. The two most important questions we will be asking about each type will be:

1 How much confidence can be placed in the results of this kind of test?
2 Exactly what line of reasoning justifies the making of decisions on the basis of such tests?

2.2 Performance-referenced testing

As we saw above, performance-referenced tests are a relatively recent development in language testing. We are going to deal with them first, however, since they are based on rather more straightforward principles; principles which they share, furthermore, with vocational and professional tests outside the field of language testing.

2.2.1 Direct testing

Let us start by distinguishing two kinds of performance:

The test performance: i.e. what the testee has to do during the test
The criterion performance: i.e. what the testee would have to do in a 'real' situation.

The relationship of the test performance to the criterion performance can be simply expressed as follows: