

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

LANGUAGE TESTING

江苏工业学院图书馆
藏书章

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

To Jone

© 1989 David Baker

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

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronically or mechanically, including photocopying, recording or any information storage or retrieval system, without either the prior permission in writing from the publisher or a licence permitting restricted copying. In the United Kingdom such licences are issued by the Copyright Licensing Agency: 33-34 Alfred Place, London WC1E 7DP.

Photoset in Linotron Ehrhardt 10/11½pt. with Unifers 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 Résumé	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 linguistics	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

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 tests 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.

Language 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 into 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:

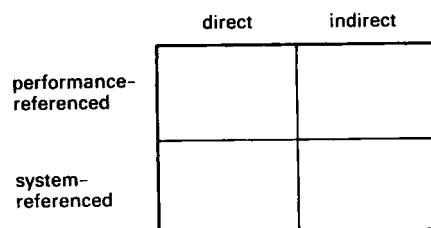


Fig 2.3

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.

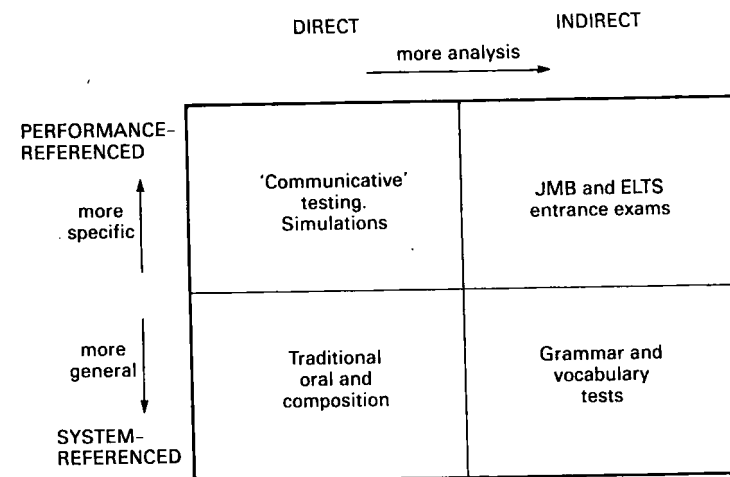


Fig 2.4

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:

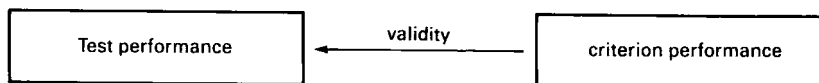


Fig 2.5

A well-known test which conforms closely to this model is the British driving test. The test performance consists of driving around and making manoeuvres for a short time. The criterion performance consists of driving around and making manoeuvres for the rest of the testee's life.

As mentioned earlier, language tests of this type are comparatively recent. Figure 2.6 shows an example from the Oxford Syndicate's preliminary test in English.

Here the criterion performance is using an English-English dictionary to resolve problems in reading. The test performance is very similar and on this basis we could be fairly confident that a candidate who performs well on the test will be able to use a dictionary of this kind effectively in the future. This illustrates a general principle of direct testing: if the test performance is sufficiently similar to the criterion performance then judgements made about the testee during the test can be considered as applicable to the criterion performance and decisions made accordingly.

Thus the driving examiner who considers that a candidate's performance during the test has been satisfactory can, with reasonable confidence, assume that his future performance will be satisfactory and therefore grant him a licence.

In discussing whether decisions can confidently be made on the basis of a test's results, we are talking, in the most general sense, about its validity. As we shall see later, several different senses and aspects of validity have been distinguished, but, following our operational definition of a test's purpose as the facilitation of decision-making, we can sketch out a basic definition in the following way: The validity of a test is the extent to which confident decisions can be made on the basis of its results.

It follows from this that the validity of a test is dependent on the purpose which it is supposed to serve. Thus a test which allows the making of one type of decision may, by this token, be invalid if its results are used as the basis of a different type of decision.

This is one way of resolving the old dispute about whether oral interviews are valid tests since shy students may do badly through no fault of their own. The usual debate on this issue revolves around whether one can separate linguistic from interpersonal skills. Approached in this way it is probably not capable of being resolved one way or the other.

If we consider, instead, the decisions that flow from the test's results then the issue is clearer: if I am selecting potential sales representatives or receptionists then the interview may be a reliable guide to the suitability of the candidates. On the other hand, if I must decide whether to reward a learner for the effort he has put into his language studies, I will have less confidence in the procedure. It should be noted that this resolution of the issue sidesteps knotty problems such as 'What is the test supposed to measure?' and 'What does it measure?' which may often be very difficult to answer. By asking instead whether the test will permit the necessary

- Remember that you may use your English-English dictionary
(You are advised to spend about 25 minutes on this question)
- 5 Using the dictionary extracts, answer the following questions.
- 5.1 The definitions of **give** and **get** (opposite) are numbered. Put the number of the correct definition for **give** and **get** as used in the following sentences, in the space provided. The first one has been done for you.
- | | | |
|-----|--|---------------|
| (a) | He has given himself to the cause. | 8 |
| (b) | This is getting very difficult. | |
| (c) | The girl gave everybody a present. | |
| (d) | I'm going to keep trying to open this bottle; I'm sure it will finally give . | |
| (e) | Did you get that cough from your sister? | |
| (f) | I just don't get it; I can never understand what he does. | |
| (g) | Did you remember to get that coat from the cleaner's? You promised to collect it yesterday. | |
| (h) | That music gives great pleasure. | |
| (i) | How much did you give for your bicycle? | |
- 5.2 Study these sentences and mark each one either correct (✓) or incorrect (x).

get /get/ v. (pres. part. getting, past part. & past tense got /got/) 1 have something: Nick's got blue eyes. 2 buy or take something: We must get some more butter. 3 fetch someone or something: Jenny will get the children from school. 4 receive something: I got a lot of presents for my birthday. 5 catch an illness: Sarah got mumps from her brother. 6 understand something: I don't get what you are saying. 7 become: I'm getting cold—please close the window. 8 come or go somewhere: When will the train get to Cambridge? 9 make someone or something move: Quick, get the children out of the burning house! get about, go or travel to many places: The old man doesn't get about much these days. get at, be able to reach or come to a place: I tried to pick the apple but I couldn't get at it. get away, leave; escape: Two tigers got away from the zoo last night. get away with, (a) do something safely, which usually brings trouble: He cheated in the exam and got away with it. (b) steal or take something: The thief got away with £5,000. get back, return: I got back from my holiday yesterday. get in, come to a place: The train got in late. get someone in, ask someone to come to the house: We got the doctor in to see our sick child. get into, put clothes on: My shoes are too small—I can't get into them. get off, (a) leave: We must get off at once or we'll be late. (b) not be seriously punished, hurt, etc.: I can't get my car to start. get together, meet; come together in a group: The whole family got together for Christmas. get up, stand up; get out of bed: It's time to get up, children! get up to, (a) do something, usually bad: I must go and see what the children are getting up to. (b) come to a place in a book, etc.: We got up to page 17 in our story today. have got to, must do something: I have got to leave soon.

give /giv/ v. (past part, given /gɪvn/, past tense gave /geɪv/) 1 hand something to someone: Mother gave me a glass of milk. 2 let someone have something: They gave us a lovely holiday. 3 pay money for goods: I gave £60 for my new watch. 4 bring a feeling, etc. to someone: The old car is giving a lot of trouble. 5 make or bring something: The sun gives light and heat. 6 send out a sound, noise, movement, etc.: Diana gave a cry when she opened the letter. 7 say that someone may have or do something: I'll give you ten minutes to change. 8 use all your time, power, etc. to do something: Schweitzer gave his life to helping sick people. 9 pass a sickness to someone else: Robert gave me his cold. 10 become weaker and less firm: The branch of the tree gave, but it did not break. give someone away (a) tell a secret about someone: I'm going to hide from my brother behind the tree—please don't give me away! (b) hand a bride to the bridegroom at a wedding.

Fig 2.6

decisions to be made with confidence, we can make a first rough estimate of the validity of a proposed procedure.

Measurement and judgement

So far we have been speaking as if a direct test were merely a matter of simulating a situation which gives the candidate a chance to show what he can do. There are, however, two other aspects of the test procedure which we have not yet mentioned: **measurement** (i.e. the assigning of a score to the performance) and **judgement** (the pass/fail decision or other recommendation). These must form part of any test procedure because at the end of the day, someone must decide 'how good' the performance was and if this was 'good enough'.

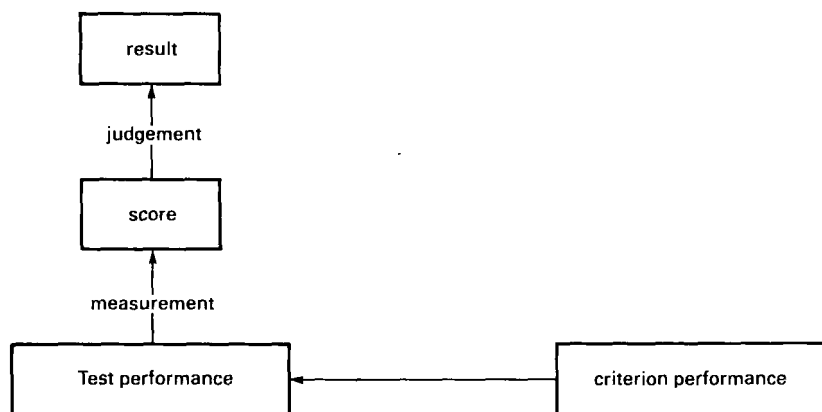


Fig 2.7

Unless the test was of an 'all or nothing' kind where nothing less than perfection is acceptable, we need some way of analysing the test performance to indicate how far it was from being perfect. In the driving test, for example, the examiner uses a checklist of sub-tasks which are ticked off as they are successfully performed by the candidate. Assessing language performance is more difficult and use is often made of a range of 'band descriptors', one of which is chosen as describing most clearly the performance being assessed, thus giving a score to the candidate. (See Chapter 6 for examples and discussion of this.)

Once the performance has been given a score, the next stage is the **judgement**, which involves deciding if the score is 'high enough'. This 'passing score' is usually decided in advance and will, of course, depend on the purpose for which the candidate is being tested. For the driving test the candidate may be allowed to perform badly on a certain number of sub-tasks without being considered a fail. The setting of this pass-mark is easier the more the test performance resembles the criterion performance. This is because it is easy to use the same criteria to judge the test performance as we would use to judge the real performance. When the two performances are rather different (i.e. the test is not very direct) it is less easy to

transfer the criteria from the real situation to the test situation and it becomes less easy to stipulate how good the test performance needs to be in order to indicate a good criterion performance.

One of the advantages of the direct type of test is that it renders the setting of the pass-mark rather easier. If we consider a test which is designed to show whether a candidate could make effective notes in a lecture, a very direct test would be to require him actually to make notes from a video-recording of a lecture. The assessor could then examine the notes and decide whether they could later be used to reconstitute the content of the lecture – this being the normal criterion of adequacy in note-taking. There would, naturally, be problems in getting exact agreement among assessors about the scores for particular performances but these can be minimized (see Chapter 6).

If, on the other hand, it was decided to use a multiple-choice listening test for this purpose (i.e. an indirect test), the setting of the pass-mark would be more problematical. Although the *measurement* would be simpler, since the multiple-choice format gives the same source whoever marks it, it would be difficult to decide, just by looking at it, what constituted a satisfactory score. The normal criteria for note-taking proficiency would be inapplicable and the pass-mark would have to be established in some indirect way. Ease of measurement is achieved at the expense of difficulties in making the judgement.

Sources of invalidity

When it is not felt that confident decisions can be made on the basis of a test then, as we have seen, the test's validity is called into question. In the case of the 'direct' test, the cause of this invalidity may be one of two things:

Firstly, the test and criterion performances may not be sufficiently similar to warrant extrapolating from one to the other, i.e. it may be doubted whether the test is in fact a direct test. This problem underlies discussions of 'authenticity' in task-based testing as we shall see later.

The second source of invalidity derives from the fact that, while life is long, tests (mercifully!) are short. It may be possible to include only a small number of features of the criterion performance in the test performance. If this part is not a representative sample of the whole performance our confidence must be weakened and the validity of the test is called into question. A driving test which only involved reversing and parking, for instance, would not be a sound basis on which to award licences.

This problem, which is usually discussed under the heading of sampling, has great importance for the kind of language tests which conform to the 'direct' model. As we have seen, the validity of this kind of test rests on the similarity between the test and criterion performance. It is the job of the test designer to set up a test situation to elicit a test performance from the candidate which will be sufficiently similar to the criterion performance. Some types of performance are easier to elicit than others, and there is a danger that the tasks chosen will be chosen for their ease of administration rather than because they are a representative sample of the criterion performance. An example of this is the much-used examiner-candidate oral interview, which is easy to set up. It is unlikely, however, that this rather unbalanced, inquisitorial speech situation will feature in much of the candidate's future performance. Of course, administrative constraints may force us to adopt such

procedures, but this should be done in the clear realization that the sampling of the criterion performance is imperfect.

Sources of unreliability

Even if the test performance is a good representative sample of the criterion performance this is not sufficient to make it a satisfactory test; as we have seen, measurement and judgement are essential stages in the administration of a test and things can also go wrong at this point. The adequacy of these aspects of the test form part of what is usually called the **reliability** of the test, i.e. the **stability** of the test as a measure. Very few candidates take the same test twice, of course, but a reliable test would be expected to give comparable results on repeated administrations. Some variations between scores would be extraneous to the test (the candidate might have been particularly tired during one administration), but other variations would be due to defects in the measuring and judging procedures of the test itself. For direct tests the chief source of this kind of unreliability is the person who measures and judges. A candidate who failed a driving test in the afternoon because of his poor reversing might have passed had he taken the test in the morning when the examiner was in a better mood. In language tests similar fluctuations occur in the severity of examiners' judgements and any test which does not have an 'objective' scoring format will have less than perfect inter-scorer reliability. (Methods of minimizing such variation are discussed in Chapter 6.) The validity of the test as a basis for decision making is, of course, dependent on the adequacy of the measuring and judging stages of the procedure as well as the sampling and selection which go on when the test is being designed. In order to be valid, then, a test must also be reliable.

Summary

So far we have been looking at tests in which there is a high degree of similarity between the test and criterion performance. We have seen that this similarity is an important part of the fundamental validity of such tests and that a rule-of-thumb definition of validity is the confidence with which we can base decisions on the results of the tests. Direct tests, in common with all other types of test, involve measurement and judgement of the candidate's performance. Measurement in direct tests tends to be problematic, while judgement is facilitated by the applicability of criteria from the 'real' performance. Inadequacy of measurement and judgement in a test contribute to unreliability which, in turn, compromises the validity of the test as a basis for decisions.

2.2.2 Indirect performance-referenced testing

Like the direct version, the indirect performance-referenced test looks ahead to a future or potential task which the candidate will or may have to carry out. In describing the indirect test we can again distinguish a test and a criterion performance. This time, however, the test criterion performances are not very similar. This is because the test performance has been derived from the criterion performance by a process of analysis and abstraction. Of course some kind of abstraction takes place even in the construction of a direct test: the checklist which the driving-test examiner uses represents a partial breakdown of driving activity into sub-tasks. In

indirect tests, however, this process of analysis and abstraction is taken much further, and results in the candidate doing things during the test which are quite different from the kind of performance that the test is designed to give information about.

As an example of this, let us consider part of a university entrance exam for post-graduate students: the criterion performance might be doing post-graduate work in an English-medium university; the test performance is reading some sentences and putting crosses in boxes on a sheet of paper. Clearly, using the criterion we adopted for direct testing, the two performances are so different as to render the test completely invalid on the face of it. And yet, tests corresponding to this description are used all over the world and an enormous number of decisions about university admissions are confidently made on the basis of such tests every year. Applying our 'decision' criterion developed earlier, it would seem that these tests are considered by those who design and use them to have a high degree of validity. How can this be, when the test and criterion performances are so very different?

It is clear that the relationship between the two performances is not a simple or direct one. In order to investigate the nature of this relationship we have to introduce two more expressions:

The criterion proficiency: i.e. what the candidate must know or be able to do in order to produce a satisfactory criterion performance,

which in this case is to follow a course of study at an English-medium university.

The test proficiency: what the candidate must know or be able to do in order to produce a satisfactory test performance,

which in this case is to put crosses in the right boxes on the paper.

How do these two constructs connect to the test and criterion performances? An illustration will help to make it clear:

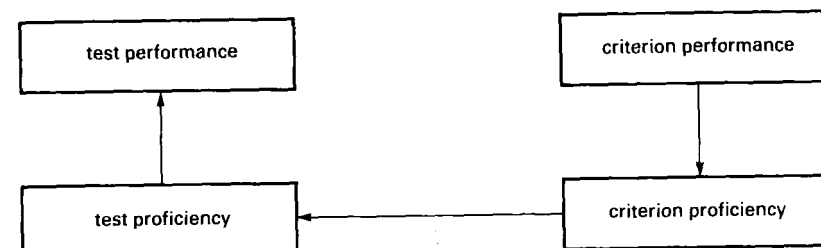


Fig 2.8

Designing indirect performance-referenced tests

The test designer, in making the test, follows a clockwise path starting with the criterion performance. Having made a description of the sort of things the testee will have to do (in our example, listening to lectures, making notes, asking questions, etc.) he then proceeds to identify what the candidate needs to know or be able to do in order to carry out the tasks he has described. This process of analysis can be done in a number of ways. We will be looking in later chapters at some of the ways of doing this analysis which have been proposed by test designers.

Having produced an inventory of the necessary proficiency, the test designer finds himself up against a similar kind of sampling problem to that faced by the designer of direct tests. If the analysis has been done thoroughly, the criterion proficiency will usually have many more aspects than can be dealt with in the course of an average language test. For instance, it might have been decided that a non-technical recognition vocabulary of four thousand items is necessary to follow a particular course. It is obviously out of the question to include all four thousand of these items in the test. A selection must be made. The resulting sub-set of the criterion proficiency constitutes the test proficiency. The task of the designer is now to produce a test which will indicate to what extent this proficiency is possessed by the candidate. This test must elicit from the candidate a test performance which can legitimately be considered as evidence for the specified test proficiency.

In this way, by passing in a clockwise sense round the diagram, we have established the relationship between the test and criterion performances. Not, as in the direct test, by examining the degree of similarity between them, because they are quite dissimilar, but rather, by specifying a number of procedures which, if legitimate, will permit judgements of the test performance to be extrapolated to the criterion performance.

As in the direct test, there are procedures for measuring and judging the test performance, although the problems tend to be slightly different in the case of indirect tests. We shall have more to say about this later on.

An example of an existing test which illustrates these design procedures is the ELTS test, produced by the British Council to enable British universities to assess the suitability of overseas applicants.

The stages leading to the construction of this test can be summarized as follows. (A far more detailed account can be found in Carroll, 1980.)

First the criterion performance is established by specifying which situations prospective students will find themselves in and what activities they will engage in in these situations. A very brief example of this specification for a Business Studies student is shown in Figure 2.10.

Next, the skills which the student will need in order to perform these activities are listed (in this case they are drawn from Munby, 1978). These constitute the criterion proficiency. The diagram shows a few of these which are relevant to Reference study – Intensive reading. A selection of these skills is made – the Test proficiency – and a test constructed to measure to what extent the candidate possesses these skills. A few items from the Study Skills module of the ELTS are shown in the diagram.

This account is, of necessity, oversimplified and does not touch on many of the issues and problems that this kind of project must deal with. It does however illustrate the sort of approach which is necessary to ensure the maximum validity of

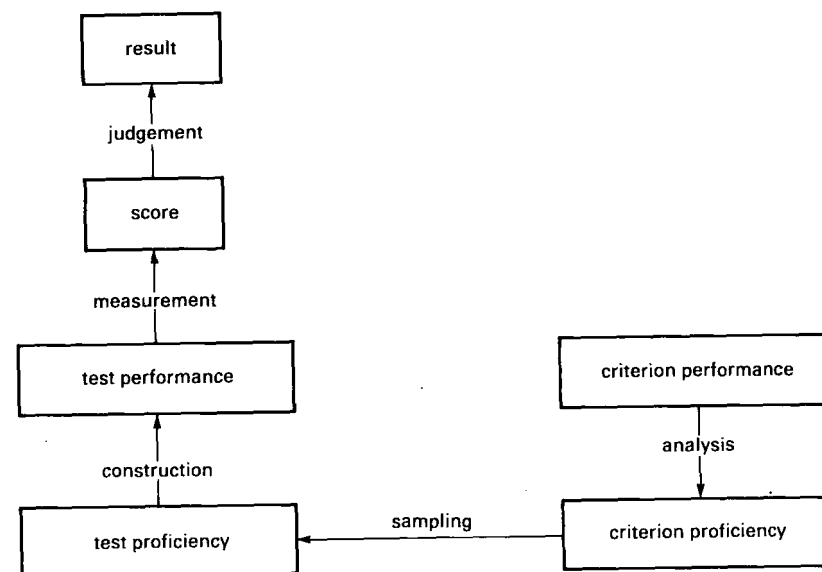


Fig 2.9

an indirect test.

Using the test

This model can also be used to make explicit the reasoning of the test-user who is basing decisions on the test. Faced with a candidate who has scored satisfactorily on the test he starts with the test performance and proceeds anti-clockwise. Since the candidate's score was satisfactory and he believes the test to be well-constructed and properly marked he has reason to believe that the candidate possesses the test proficiency. As he believes the test proficiency to be a representative sample of the criterion proficiency he can go further and deduce that the candidate also possesses the criterion proficiency. Since he believes the criterion proficiency specification to be based on a sound analysis of the criterion performance he has reason to believe that the candidate will be capable of producing this performance. In the case of our example he can therefore offer him a place on the course.

Sources of invalidity in indirect performance-referenced tests

One thing which is clear from the above is that our test user is having to place quite a lot of faith in several different aspects of the test procedure. Only one step in the test construction needs to be faulty for the legitimacy of reasoning from test to criterion performance to be seriously undermined. As we established earlier, the confidence with which we can base decisions on test results is a measure of the basic validity of the test. The kinds of problem which can creep in at each stage of the procedure represent, therefore, a number of sources of invalidity which may affect this kind of indirect test. Let us examine them.

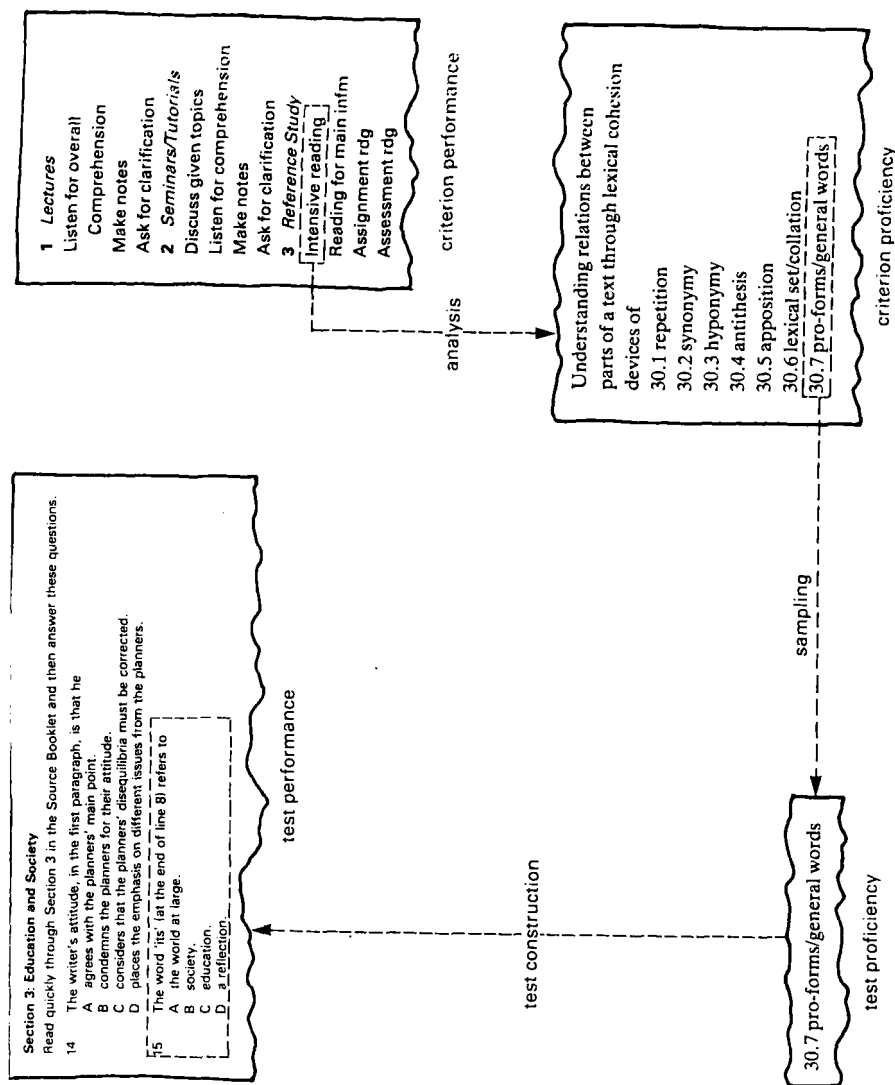


Fig 2.10

(a) Faulty analysis

In criticizing a given test it may be questioned whether the criterion proficiency is really an adequate analysis of the knowledge and skills necessary for the criterion performance. In other words the legitimacy of the passage from criterion performance to proficiency may be placed in doubt. It may be criticized as incomplete as, for instance, a university entrance test based solely on knowledge of vocabulary. Alternatively the whole basis of the analysis may be called into question and it may seriously be doubted whether the analysis of the criterion performance can legitimately be carried out in the terms proposed. In traditional testing terminology, the satisfactoriness of this analysis is usually discussed under the heading of 'construct validity'.

(b) Bad sampling

Assuming that our specification of the criterion proficiency is satisfactory, we still have to decide which parts of it will form the content of the test. The same sampling problems arise as for direct testing, accompanied by the same temptations: some aspects of the criterion proficiency will be easier to test than others. It is very easy to take as our test proficiency only those aspects which will give us fewer problems of administration, even though they may not be a good sample. The road to mediocre testing is paved with bad samples. This aspect of the test's adequacy is traditionally termed its 'content validity'.

(c) Bad construction

We now have to look at problems which may arise in the passage from the test proficiency to the test performance. This is the last stage in the design procedure and concerns the practical construction and administration of the test. The test writer's job is to produce a task which will enable the testee to demonstrate if, or to what extent, he possesses the test proficiency. Clearly, anything which prevents a candidate who has the relevant proficiency from producing a satisfactory performance renders the test invalid. Such influences include bad instructions, insufficient time and 'trick' questions.

Conversely, anything which permits a satisfactory performance by a candidate who does not possess the relevant proficiency also makes the test invalid. This category includes all kinds of cheating and things which facilitate guessing.

These problems, together with inadequacies in the measurement and judgement of the test performance (examined in the next section) constitute sources of **unreliability**. That is, they introduce variations in the results of the test which do not reflect the candidate's test proficiency. As in the direct test, the unreliability of the test as a measure compromises its use for decision-making and thus its validity.

(d) Faulty measurement and judgement

Unlike the direct test, where it may be difficult to measure a performance reliably, this is rarely a problem in indirect tests. The analysis of the criterion performance permits the construction of a test which consists of a number of sub-tasks and items, and the result of the test can usually be expressed numerically without any trouble. Where the test consists of closed or 'objective' item-types (such as multiple-choice), variations due to inter-scorer unreliability can be eliminated. The problem with indirect tests lies in the making of the judgement. What is the pass-mark? What