Testing Spoken Language

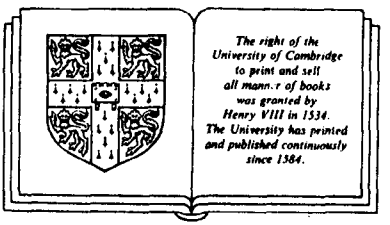
A handbook of oral testing techniques

Nic Underhill

Testing Spoken Language

A handbook of oral testing techniques

Nic Underhill



Cambridge University Press
Cambridge
London New York New Rochelle
Melbourne Sydney

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
32 East 57th Street, New York, NY 10022, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1987

Printed in Great Britain at the University Press, Cambridge

Library of Congress cataloguing in publication data

Underhill, Nic.
Testing spoken language.

(Cambridge handbooks for language teachers)
Bibliography
Includes index.

1. Language and languages – Ability testing.
2. Language and languages – Examinations. 3. Oral communication – Ability testing. 4. Oral communication – Examinations. I. Title. II. Series.
P53.6.U5 1987 001.54'2'076 86–31052

British Library cataloguing in publication data

Underhill, Nic

Testing spoken language: a handbook of oral testing techniques. – (Cambridge handbooks for language teachers)
1. Language and languages – Study and teaching – Great Britain. 2. Language and languages – Ability testing
I. Title
418'.0076 P57.G7

ISBN 0 521 32131 X hard covers
ISBN 0 521 31276 0 paperback

WD

Acknowledgements

My thanks are due to Clive Jaques, John de Jong, Mike Levy, Keith Morrow, Don Porter, Bill Shephard, Norman Whitney, Protase E. Woodford and my colleagues at International Language Centres, for their ideas, encouragement and help; and to Ush for her invaluable common sense.

Contents

Acknowledgements vii

Introduction 1

1 Aims and resources 11
1.1 Aims 12
1.2 Resources 15
1.3 Needs 18
1.4 Expectations 19

2 Test types 22
2.1 Self-assessment 22
2.2 Teacher assessment 27
2.3 Who does the learner speak to? 28
2.4 The direct interview type 31
2.5 The pre-arranged information gap 32
2.6 Tests where the learner prepares in advance 33
2.7 Mechanical/entirely predictable tests 33
2.8 Recording oral tests 34
2.9 Using the telephone 36
2.10 Sequencing test techniques 37
2.11 Timing and location of tests 39
2.12 On being friendly 42

3 Elicitation techniques 44
3.1 Discussion/conversation 45
3.2 Oral report 47
3.3 Learner–learner joint discussion/decision making 49
3.4 Role-play 51
3.5 Interview 54
3.6 Learner–learner description and re-creation 56
3.7 Form-filling 58
3.8 Making appropriate responses 59
3.9 Question and answer 61
3.10 Reading blank dialogue 64
3.11 Using a picture or picture story 66

3.12 Giving instructions/description/explanation 69
3.13 Precis or re-tell story or text from aural stimulus 71
3.14 Re-telling a story from written stimulus 73
3.15 Reading aloud 76
3.16 Translating/interpreting 79
3.17 Sentence completion from aural or written stimulus 81
3.18 Sentence correction 84
3.19 Sentence transformation 84
3.20 Sentence repetition 86

4 Marking systems 88
4.1 The number of assessors 89
4.2 The selection and training of assessors 90
4.3 Marking recorded oral tests 92
4.4 Marking keys or marking protocols 94
4.5 Mark categories 95
4.6 Weighting 97
4.7 Rating scales 98
4.8 Impression marking 100
4.9 Additive marking 101
4.10 Subtractive marking 102

5 Test evaluation 104
5.1 Face validity 105
5.2 Content validity 106
5.3 Construct validity 106
5.4 Reliability 107
5.5 Concurrent validity 107
5.6 Predictive validity 108

Appendix I Three public oral tests 109

Appendix II Bibliography and further reading 112

Index 116

Introduction

This introduction is divided into six sections:

Section 1 describes who the book is intended for;
Section 2 presents a model;
Section 3 asks why such a book is necessary at all;
Section 4 summarises the themes on which the book is based;
Section 5 is a glossary that defines some testing terminology for the purposes of this book;
Section 6 explains the order and content of the chapters.

WHO IS THIS BOOK FOR?

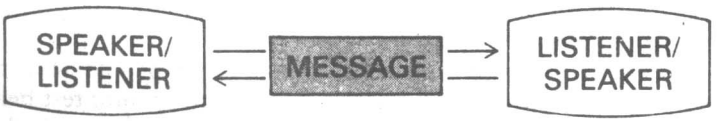
This handbook is intended for teachers and other people who are interested in the use of oral tests of language ability; an oral test being defined as a test in which a person is encouraged to speak, and is then assessed on the basis of that speech. The book is aimed at any kind of language teaching programme where it is desired to produce an oral test that will fit in with the teaching programme and that will be designed by, or in full consultation with, the teaching staff themselves. The sequence of this book follows the order in which a new test programme would logically be carried out; starting with preliminary questions about needs and resources, then presenting different oral test types and tasks to choose from, and finally discussing the marking system and evaluation of the test in practice.

Answers to questions in chapter 1 will affect choices in chapter 2, which will in turn constrain the choice of testing and marking techniques in chapters 3 and 4. Although they are presented in this logical sequence, these different steps in the test programme are not distinct operations to be carried out one at a time, independently of each other. To produce the 'best' test, you will need to consider how decisions made in one area will affect your freedom of choice in another.

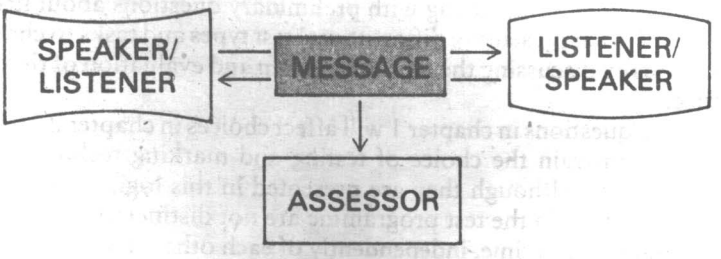
The book does not assume any knowledge of language testing as a specialist discipline and is written for practising language teachers. It deplores the cult of the language testing expert. All the examples of test types are in English, but they could equally well be applied to oral tests in other languages.

A MODEL

The model below is sometimes used to identify the different components involved in communication by speech. The arrows indicate the direction of speech. They point in both directions; at one moment, one person is listening to the other person speaking, and the next moment, the roles may be reversed. The speaker becomes the listener, and the listener becomes the speaker. These switches from one role to another often happen very fast in conversation. Speech is normally a two-way system of communication: situations where only one person speaks and others only listen, such as an academic lecture or a political address, are comparatively rare. This feature of interactive role-switching distinguishes good oral tests from other language tests; listening, reading or writing tests which present a set of questions and elicit a set of answers are clearly not interactive in this way.

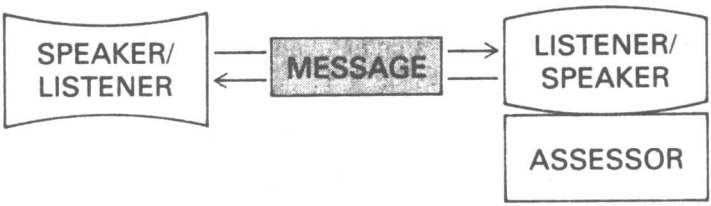


With one addition, the same model can be used to represent the oral test situation. As well as a person who speaks and a person who listens, in an oral test we need somebody to assess that speech. It is this process of assessment that turns it into a test.



In an oral test, you do not need to have three different people, one for each role. Chapter 2 uses variations of this model to describe and compare different test types. Some types of oral test have more than three people, some have fewer: self-assessment, for example (see 2.1) needs only one person. The most common type of oral interview involves two people, the learner and a person who is both listener and assessor. (See below.)

This test type, for example, is economical but it does require somebody to carry out two roles at the same time, and this can be difficult to do.



WHY A BOOK SPECIFICALLY ON ORAL TESTING?

Oral tests are qualitatively different from other kinds of tests. They do not easily fit the conventional assumptions about people and testing, which are examined below. There is a lot of interest now in oral testing, partly because teaching is more than ever directed towards the speaking and listening skills, particularly in the early stages. Naturally, this should be reflected in the testing. In order to free oral tests from the burden of conventional language testing wisdom, they should be considered as a class of their own, and that is the purpose of this book.

Why have oral tests generally received little attention? Many books have been written about language testing. They follow the changing fashions of language teaching, but they usually make the same basic assumptions about the nature of language testing. Generally, little space is devoted to oral testing compared to testing the other skills. This is partly because of the difficulty of treating oral tests in the same way as other more conventional tests.

The most striking assumption in these books is that it is the test itself that is important, while the human beings who take the test, and those who mark it, seem less important. Tests are seen as objects with an identity and a purpose of their own – you can see many references in the literature to the 'test instrument' – while the learners are often remote figures who are only of interest for their reactions to this instrument (the usual term in testing research for test-takers is 'subjects'!). The number of possible ways they can react is strictly limited. Similarly, the preference for easily marked, multiple-choice or limited-response tests reduces markers to the role of machines. Because of this, learners do not *enjoy* taking tests, and teachers do not enjoy marking them.

In a genuine oral test, this order of priorities is reversed. Real people meet face to face, and talk to each other. The 'test' may not even exist, in the same way that a written test does, on paper. It is the people and what passes between them that are important, and the test instrument is secondary. In fact, with a technique like an oral interview, it becomes impossible to talk about the 'test' independently of the people involved in it because it doesn't have a separate existence as a set of questions on paper.

It follows that oral tests must be designed around the people who are going to be involved. This is a human approach; we want to encourage people to talk to each other as naturally as possible. The people, not the test instrument, are our first concern.

Where do these conventional assumptions about testing come from?

The most important influence on the development of language testing has been the heritage of psychometrics, in particular intelligence testing. 'Psychometry' is the measurement of the human mind, and in the first half of this century a lot of time and effort was devoted to proving that there was a single measurable attribute called general intelligence – the 'g' factor – which we all possessed in different quantities. The 'g' factor was said to be related to parental and racial hereditary traits and to social background. Virtually all the statistical techniques used in language testing today were developed to a high degree of sophistication fifty years ago or more in the drive to prove this biological basis of intelligence. These techniques in current use in testing include correlation, regression and the analysis of variance, as well as item analysis, item discrimination, and various test reliability formulae.

The fact that the statistics actually proved nothing at all did not lead to the assumptions being challenged; it merely meant that ever more complex statistics were developed that *would* prove the initial assumptions of the researcher. These 'sophisticated statistics poised on testing techniques of rustic simplicity' (see Appendix II) became incomprehensible to the layman, and encouraged by the experts, led to an even stronger belief in the invincibility of statistical methods. This invincibility is so strong that it can easily withstand cases of scientific fraud, such as Cyril Burt's deliberate falsification of his data on twins in the 1950s.

This sounds like a remote chapter of history, interesting but irrelevant to the subject of language testing. However, it is relevant, because the same assumptions of mechanical measurability and statistical invincibility exert a strong influence on our attitudes today. In the late 1970s, the field of language testing research was revolutionised by the use of a sophisticated new technique called factor analysis, which takes the scores obtained by the same people on a number of different tests and extracts a number of different *factors*, which are supposed to relate to fundamental mental abilities.

In particular, factor analysis was supposed to prove the existence of a single factor of general language competence, which is related to performance on different language tests to a greater or lesser extent. (This, too, was affectionately known as the 'g' factor.) The implication was that all tests, including oral tests, were only partial reflections of this basic underlying language competence. Fashions change, however, and now, only a

few years later, the Unitary Competence Hypothesis, as it was called, has been largely forgotten.

Was factor analysis a 'new' technique? Not at all; it was invented in the first years of this century, and was used in exactly the same way to support the argument that there was an underlying general intelligence trait and that different tests merely reflected this more or less strongly. The issue was neither proved nor disproved convincingly; it rumbles on to this day. The Unitary Competence Hypothesis has suffered the same fate. The point is not that statistics are always wrong; as sets of figures, they are inherently neutral. But they must be treated with caution; it is easy to believe that they prove or support a theory when in fact they don't.

Psychometrics wanted to be a science. Those aspects of human behaviour that could be predicted and measured were emphasised; those aspects that were unpredictable or inconsistent were ignored. The test techniques developed then, particularly the multiple-choice format so familiar to us now, offered the learner no opportunity to behave as an individual. The statistics developed then and used today interpret human behaviour as basically similar, and conceal any form of self-expression. Such behaviour was described as 'variance', and a lot of effort was put into reducing the amount of variance a test produces.

Language teaching inherited this belief that language proficiency, like general intelligence, was a single, underlying mental factor, and that therefore the measurement of language by a test was essentially the same process as the measurement of length by a tape-measure or temperature by a thermometer. The assumption carried over from intelligence testing was that the language proficiency of an individual is at a single point on a linear scale, and that this point can be determined by an objective test instrument. All you have to do is develop the right instrument.

Language testing has developed enormously in recent years and has absorbed many other influences of a more pragmatic nature. We no longer believe that there is a single scale of language proficiency. Anyone with any experience of oral testing in particular will know that oral ability cannot be forced into such a mould. But the criteria we use for evaluating tests still favour the statistical assumptions of the mental testing heritage, and the result is a strong bias towards mechanical tests and against the human face of oral tests.

This book is an attempt to redress that imbalance. It recognises that oral tests, because they involve a subjective judgement by one person of another, are likely to be less reliable; but it suggests that the human aspect of that judgement is precisely what makes them valuable and desirable. When we test a person's ability to perform in a foreign language, we want to know how well they can communicate with other people, not with an artificially-constructed object called a language test. If we are willing to question the assumption that statistics must come first, we can then start

from the idea of the best possible test for a particular context and work to strike the right balance between the human and the statistical forces.

SUMMARY OF THEMES

1. *You need full local knowledge.* Tests are not inherently good or bad, valid or invalid; they become so when they are applied to a particular situation. You cannot say how good a hand-tool is unless you know exactly what it is used for; similarly, you can only evaluate a test in a specific context.
2. *You need to design the test as a whole.* Following on from a full awareness of the local conditions, an oral test must be conceived as an integral whole, and usually as a natural complement to the teaching programme. It's no good asking one person to draw up the aims, another to decide on the test techniques, and a third to design the marking system. The entire procedure should appear as a single and consistent entity to testers and learners alike.
3. *You need a human approach.* Oral tests must treat people as human beings. In small ways, as well as in the design of the test procedure in general, we can make taking a test challenging, instructive and even an enjoyable experience. There is a good practical reason for this, not just that it is nice to be nice; if you treat people in as friendly and human a way as possible they will tend to respond in kind, and you are going to get a much more accurate picture of their oral ability.
4. *You need to find a suitable balance.* The planning and execution of an oral test involves making positive compromises between different forces, for example, between communicative and structural aims, between impression and analytic marking systems, and between highly reliable and highly valid techniques.
5. *You need to adapt and improve.* At the same time, the balance is dynamic – no test procedure is sacred. Test evaluation is not something you do once, then sit back and relax; it is a continuous process. The best test reflects any changes in the aims of the programme or the needs of the learners. There are so many factors that have to be considered in the design of a test procedure that it would be surprising if circumstances did not change from time to time.

GLOSSARY

The words described below are often used in testing, but with different meanings by different writers. This glossary is not intended to provide universal definitions that will work in every context; it gives working definitions of testing terms used frequently in this book, in order to help the reader understand the author's meaning.

Oral test An oral test is a repeatable procedure in which a learner speaks, and is assessed on the basis of what he says. It can be used alone or combined with tests of other skills.

Learner A learner is a person who takes a test in a foreign language. This is preferred to ***student***, as a person taking a test may not be a student at the time; we are all learners, whether or not we are students. It is also a better label than ***testee*** or ***subject***, both of which are unattractive and have connotations of the laboratory animal under experimental observation.

Interviewer An interviewer is a person who talks to a learner in an oral test and controls to a greater or lesser extent the direction and topic of the conversation. While exercising this control, she may nonetheless yield the initiative to the learner to redirect the discussion to another area. An interviewer also takes the role of the assessor, or one of the assessors. As there are so many possible variations in the roles of interviewer, interlocutor and assessor, interviewer is used as a cover-all term in this book for the person who conducts the test and deals directly with the learner.

Interlocutor Some oral tests have a person whose job is to help the learner to speak, but who is not required to assess him. An interlocutor is a person who talks with a learner in an oral test, and whose specific aim is to encourage the learner to display, to the assessor, his oral fluency in the best way possible. An interlocutor is not an assessor. She may well be known to the learner, for example, as a teacher.

Assessor An assessor is a person who listens to a learner speaking in an oral test and makes an evaluative judgement on what she hears. The assessor will be aided by pre-defined guidelines such as rating scales (see 4.7), which give considerable help in making these judgements. Ultimately, the decision is a subjective one, which is to say that it is a human one made on the basis of judgement, intuition and experience. Having more than one assessor usually means a more reliable judgement (see 4.1).

Marker This term is reserved for someone who is not present at the test itself but later awards marks to the learner on the basis of an audio or video tape recording. This may be a routine part of the marking system (see 4.3, second marking) or it may be an occasional exercise for the purposes of ***moderation***. This definition of marker reflects its use for people who correct other types of language tests; they mark the papers but they never meet the individuals who wrote them. Assessors do meet the individuals; this is one reason why assessing is different from marking and why oral tests differ from other tests.

Rater The term is used in this book as a synonym for marker. An *examiner* is a person who marks or assesses performance in stan-

dardised and large-scale language test batteries, usually containing tests of several different kinds. While an oral test may well be included in such batteries, such examinations are outside the scope of this book.

Communicative When a learner says something that is relevant and true (for himself at least), to someone else who is interested and has not heard it before (from that speaker, at least), then that act of speech is communicative.

Any definition of this term will be incomplete, and invite correction, or at least addition. For example, no mention is made here of the continuous interactive exchange between speakers which is characteristic of normal conversation. This is a minimal working definition. This book describes techniques and procedures as more communicative or less communicative rather than the black-and-white terms communicative or non-communicative. The term is used in this book in a neutral sense – a more communicative test is not necessarily a better one. A number of less communicative techniques are included in the following chapters; there are many circumstances in which it is desirable to use these as well as, or instead of, more communicative techniques.

Authentic An authentic task is one which resembles very closely something which we actually do in everyday life. To engage in free conversation is an authentic task; to transform sentences from active into passive or present to past is not authentic. Note that this is not the same thing as communicative. Copying out a shopping-list or an address is authentic but not communicative; and there are many communicative exercises, for example of the 'information gap' kind, which are not authentic.

Objective An objective test is one in which there is a single correct answer, or a very small number of possible correct answers, for each question. The marker only has to decide whether a learner has given the right answer to each question, and she is not required to exercise any personal judgement. In theory, an objective test could be marked by a machine; sometimes they are.

This is the central meaning of *objective*; the opposite is *subjective*, and both are neutral terms for particular types of tests; they do not have any positive or negative connotations. Some situations would call for one type, some for the other type. Perhaps both can be used.

Unfortunately, they have also acquired connotative values which suggest that a subjective test is necessarily bad and an objective test is always good. This will tend to be true if you are very concerned with statistical reliability (see 5.4); but if you are very concerned with communicative validity, then the reverse is true. Objective tests are easier to mark but they are almost always less realistic.

Stimulus A stimulus is something that is intended to encourage the learner to speak, usually by providing a subject to talk about. It might

be a picture, a text, an object or a particular topic. Its use in this book does not have any of the technical meaning of the automatic stimulus/response theory of behaviourist psychology.

Validity As a general term, does the test measure what it's supposed to? Having specified the aims of a test at the outset, the purpose of validation is to find out if in fact it meets those aims. Different forms of validity are discussed in more detail in chapter 5.

Reliability Does the test give consistent results? If the same learners are tested on two or three occasions, do they get the same score each time?

Evaluate To evaluate a test is to find out how well it is working, in the widest sense. Is it valid? Is it reliable? Does it take too long? Are the learners and the assessors happy with it? Does it meet the specifications?

Moderate Moderate has a more restricted meaning, concerning test reliability. To moderate a test is specifically to compare the way different assessors award marks, and to take steps to reduce any discrepancies.

Best test This is the test most suitable for the particular situation in which it will be used. Although every test procedure has its own advantages and disadvantages, it is only a good test or a bad test in a particular context. If somebody asked you what the best car is, you would probably say that it depends on what you want to use it for; a Jeep is better than a Jaguar for driving over rough ground, and a Mini is better than a Cadillac for driving round a small, crowded town. Because the best test is context-specific, it can only be designed for, and produced in, that context, and no amount of expert authority will make the best test in one place automatically the best test in another.

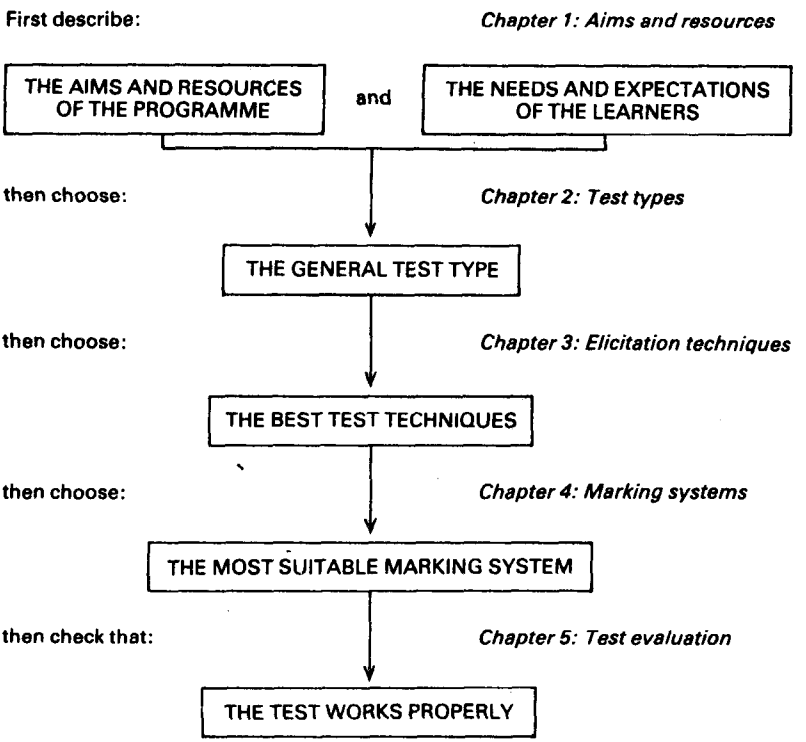
He and *she* are used throughout the book as referring to the learner and the tester/interviewer/assessor respectively. By this means, the presence and involvement of people of both sexes may be implied, while the potential ambiguities of pronominal reference are avoided.

At the end of the book are two appendices. Appendix I describes three oral tests that form part of publicly available test batteries, to show how some of the testing and marking techniques described in this book are used in practice.

Appendix II contains a number of bibliographical references and suggestions for further reading. In the interests of readability, and to further the aim of demystifying testing by removing the shadow of the expert, the main text contains no references at all.

THE STRUCTURE OF THE BOOK

The order of the chapters follows the steps in the development of an oral test procedure.



1 Aims and resources

This chapter asks a number of questions about the general background in which the oral test is to be designed and used. Sections 1 and 2 ask about the institutional aims and resources, and sections 3 and 4 ask about individual needs and expectations. There are many different types of oral test and marking systems, and the answers to the questions posed in this chapter will make it easier to select the most suitable techniques from the following chapters.

The first section asks the question we should ask before we start any new project: why exactly are we doing it? If you know precisely what the aims of the project are before you begin, it will be much easier to take the right decisions later on. When you reach the evaluation stage (see chapter 5) it will also be easier to decide if you have achieved those aims. The question is not always as easy to answer as it sounds; many language tests are given because it is the accepted practice to give language tests as part of a teaching programme, without setting out clear aims.

A written multiple-choice test is usually held in scheduled lesson time in an ordinary classroom, without the need for any special arrangements. Oral tests, on the other hand, can often be more difficult to design, administer and mark. It is all the more important, therefore, to make sure that you know in detail the purpose of the test before you begin and so prevent resources being wasted on producing a test that is unsuitable.

Available resources are the subject of the second section. When you know what the aims are, the next question is: what resources do we have to help us achieve those aims? These resources include people, time, space and equipment. Every oral test technique requires slightly different resources, and the best test will consist of techniques which match the resources available. For example, choosing a test that requires a brief interview with a native-speaker if you have lots of time but only one native-speaker available, would be an inefficient use of resources.

The third and fourth sections ask about the needs and expectations of the learner. The person who takes a test is the immediate consumer of the product; when manufacturers are designing a new chocolate bar or style of furniture they go to great lengths to find out what their potential customers need or want. In the history of language testing, on the other hand, we have often managed to ignore the point of view of the test-takers altogether. A good oral test allows learners to be treated, and to behave,