Edited by ALAN DAVIES



Language Testing Symposium

A Psycholinguistic Approach





Language Testing Symposium

A Psycholinguistic Approach

Edited by

ALAN DAVIES

London
OXFORD UNIVERSITY PRESS

Oxford University Press, Ely House, London W. I London oxford glascow new york
Toronto melbourne wellington cape town
IBADAN NAIROBI DAR ES SALAAM LUSAKA ADDIS ABABA
KUALA LUMPUR SINGAPORE JAKARTA HONG KONG TOKYO
DELHI BOMBAY CALCUITA MADRAS KARACHI

© Oxford University Press, 1968

ISBN 0 19 437033X

First published 1968 Third impression 1976

PRINTED IN GREAT BRITAIN BY HEADLEY BROTHERS LTD 109 KINGSWAY LONDON WC2B 6PX AND ASHFORD KENT

ACKNOWLEDGEMENTS

Most of the contributions to this volume now appear in print for the first time. The Symposium was planned as a unity, as an approach to Language Testing from a number of related points of view. This plan involved commissioning papers from those with wide experience in the area. The Editor wishes to thank all those who have contributed. In addition the Editor would like to thank Mr R. Mackin and Dr A. E. G. Pilliner for advice and encouragement, Mr S. P. Corder for suggestions regarding the Introduction, the OUP for patience and Mrs Ethel Jack and Mrs Pat Bertram for help in typing.

The following have given permission for copyright material to be printed here: McGraw-Hill Book Company, New York, for material published in Chapter 11 of Trends in Language Teaching (edited by A. Valdman), by P. Pimsleur. Copyright © 1966 by McGraw-Hill, Inc.; the West African Examinations Council, for parts of the Report on Oral English Testing in West Africa, 1965, by A. Davies; the West African Examinations Council and the African Universities Press for Chapters 7, 13, and 16 of the Report on English Language Examining, 1963, by D. W. Grieve.

All references to books and articles throughout the text are indicated by author and date and will be found in the General References at the end of the book. In addition a Test List is given of Language Tests in print.

Spelling variants

No attempt has been made to make all contributions conform to one spelling convention. Where there are British and American differences these have been retained.

Hyphenization is even more arbitrary. Again individual uses have been retained.

CONTRIBUTORS

KEITH BROWN

taught for British Council in Uganda, lecturer in English, University College of Cape Coast, Ghana. At present lecturer in Linguistics, University of Edinburgh.

J. B. CARROLL

taught Psychology in universities; Professor of Educational Psychology at Harvard; now joined Educational Testing Service, Princeton, New Jersey, as Senior Research Psychologist. Author of The Study of Language, Language and Thought, Language, Thought and Reality, and with Sapon Modern Language Aptitude Test, etc.

ALAN DAVIES

taught English in secondary schools in England and East Africa. At present lecturer in Dept. of Applied Linguistics, University of Edinburgh.

D. W. GRIEVE

taught English in West Africa; assistant registrar of West African Examinations Council; now Senior Lecturer in English in the University of Ghana. Author of English Language Examining and many textbooks in English as a Second Language.

D. P. HARRIS

taught English Language in universities; worked on English Language Test construction at Michigan, Georgetown, and Educational Testing Service; first Director of the Program for the Testing of English as a Foreign Language (TOEFL); at present Director of the American Language Institute and Professor of Linguistics at Georgetown University, Washington, D.C. Author of several textbooks for students of English as a Foreign Language.

ELISABETH INGRAM

is a psychologist. She has been a lecturer in the Dept. of Applied Linguistics, University of Edinburgh, since 1957; and has written articles on language learning and a Test of English Proficiency.

G. E. PERREN

trained teachers of English at all levels in East Africa; formerly Director of English-Teaching Information Centre, British Council, London. Since 1966 first Director of Centre for Information on Language Teaching, London.

A. E. G. PILLINER

taught Science and Maths. in secondary schools; since 1949 Lecturer, then Senior Lecturer in Education and Director, Godfrey Thompson Unit for Educational Research, University of Edinburgh.

Author of articles in statistical and educational journals and part-author of Moray House tests.

PAUL PIMSLEUR

taught Modern Languages in universities; Professor of Modern Languages and Education and Director of the Listening Center, Ohio State University, since 1961; author of articles on language learning, especially under-achievement, of Language Aptitude Battery and of Pimsleur Proficiency Tests in foreign languages.

ANDREW WILKINSON

taught English in secondary schools in England; Lecturer, now Senior Lecturer, in Education, University of Birmingham. Director of current research project into Oracy. Author of Spoken English, articles on English teaching, etc. Editor, Educational Review and English in Education.

CONTENTS

	Acknowledgements	v i
	Contributors	vii
I	Introduction	I
2	Subjective and Objective Testing A. E. G. Pilliner	19
3	The Linguistics of Language Testing D. P. Harris	36
4	The Psychology of Language Testing J. B. Carroll	46
5	Attainment and Diagnostic Testing Elisabeth Ingram	70
6	Language Aptitude Testing Paul Pimsleur	98
7	Testing Spoken Language: Some Unsolved Problems G. E. Perren	107
8	The Testing of Oracy Andrew Wilkinson	117
9	Tests in Education D. W. Grieve	133
10	Oral English Testing in West Africa Alan Davies	151
II	Intelligibility Keith Brown	180
	Appendix: Item Analysis Elisabeth Ingram	192
	References	204
	Language Tests	208
	Index of Names	211
	Index of Subjects	212

Introduction

I. LANGUAGE-LEARNING VIEWS AND THEIR INFLUENCE ON LANGUAGE TESTING

Language Testing involves both linguistics and psychology because it is concerned with language and with learning. It is also experimental because it sets up learning tasks in order to study behaviour; and also evaluative because it makes use of statistical techniques in order to study that behaviour. There are, therefore, three strands in language testing: language, learning and evaluation.

It is not only in language testing that these three strands meet: the whole of Psycholinguistics is concerned with the relationship between language and learning and makes use of statistics like any other experimental science. It is difficult, therefore, to understand how language could be considered separate from learning. However, even today many do not regard language and learning as inseparable. Those who do not will probably take some account of learning but will be mainly interested in the language apart from the learner. By 'learner' here is meant not only the Foreign Language (LF) and Second Language (L2) learner, but the Mother tongue or First Language (L1) learner too. It is possible in considering the area of language teaching and testing in which linguistics, psychology and evaluation meet to disentangle three main positions that have been taken up over the last fifty years. The position taken up has two main effects on the teaching and the testing: on the content of the teaching, i.e. what is taught, and on the concern for the needs of the learner.

The first position is taken up by those we may call the strict separatists. They are not really interested in the learner at all. They see language as a thing in itself: for them 'real' language exists in texts; there is no spoken language; learning is a matter of gobbets like dates in history. Teaching is by the grammar-translation method and what the student is expected to do is to translate, if possible, in a 'stylistic' manner. The texts where the 'real' language is held to reside most truly are literary ones and so literature is emphasized in the teaching. No consideration is given to the needs of the learner because the point of it all is essentially one of mind-training, of transfer of training. Testing follows the teaching

and is mostly concerned with setting translation tasks. Now it will be clear that the position caricatured here is the well-known classical language one which has carried over into both LF and L1 teaching. This is not to suggest that all present-day teaching of the classics follows this pattern. The associated linguistic theory is the traditionalist one, and since the learner is not taken into account, there is no theory of psychology involved.

The effect of these assumptions, held by those who take up this position, on the content of teaching material is that written texts from literature, narrowly conceived, are taken to be the language. The needs of the learner are not considered except that some attempt is made to sequence the teaching material so that the beginner starts with the easiest material and progresses to the more difficult as he becomes more fluent. The grounds on which such sequencing is done seem to be entirely subjective. Since the language is viewed as a whole there is equal stress on the past and on the present so that texts for teaching and for testing are historical as often as contemporary. In the early stages, of course, as well as the attempt to sequence material for the learner, there is also an attempt to teach him. But the language, for those who take up this position, is seen as a thing and so it is taught as if it were a series of facts. There is then, no connexion with L1 learning, but if there were, the assumption would be that the L₁ learner of Latin starts off at about 18 months running through the Mensa paradigm and parsing Balbus murum aedificavit.

The second position is taken up by those we shall call the separate but equal supporters. They see language as a machine which acts as a stimulus-response mechanism. Learning for holders of this position in its extreme form, is always shaped and consists of conditioned responses to the environment. Language itself, as a machine, has no direct connexion with the environment; it is part of human behaviour. It may be regarded as a thing but it is still separate and complete in itself. Since language is separated from the environment in this way there is little recourse to meaning: there is little need felt to link language in any meaningful way with the world to which it refers. Linguistic analysis proceeds at least theoretically from sound to sentence (cf. Harris, 1951) with the unnecessary but frequently made conclusion that learning itself proceeds in the same way.

Teaching for those who hold this position must analyse language into its parts or, as is frequently said, structures. It is strange how often this term is used in isolation to refer to grammatical structures, ignoring the fact that the other language levels, e.g. phonology, semantics, have their structures too. Such an analysis into parts is sometimes known as an Immediate Constituent (IC) analysis, and such an approach taxonomic. Teaching thus begins with the smallest bits and deliberately

works from sound to sentence, making use, in a laudable enough way, of all kinds of teaching aids and machinery. It may now be time to acknowledge that the hardware of the Audio-lingual, language laboratory, tape-recorder approach has given way to the software of Programmed Learning which actually underlay it all along. That is to say that the Audio-lingual approach at its best was essentially a Programmed Learning (PL) approach. Since the assumption of the classical PL approach has been of the Stimulus-Response kind the influence on the teaching has been to make the materials take the form of the exercise-drill type. This has been the effect on teaching content. Learning, it is assumed, takes place by generalization, by analogy on the basis of the learnt response. This position which we have labelled the separate but equal one is, in linguistic terms, structuralist, and in psychological terms, behaviourist. Testing inevitably follows on by providing drills for completion, gaps for filling and multiple choice answers for selection.

Those who take up the first position we considered set tests much as they set daily exercises; passages for translation may be regarded as a form of open-ended question. There is for this point of view no real distinction between tests and exercises so that instead of saying, as we have just done, that tests are like daily exercises, we might equally well say that daily exercises are like tests; that all teaching is doing is testing. Those who have taken up the second position have mechanized their testing so that after construction a test becomes automatic. What is more, a test in this tradition will provide information as to learning in this tradition because trouble is taken to select suitable teaching material and sequence it (cf. Mackey, 1965); such a test will say whether a student has learnt by such a method, it will say nothing about whether the student knows the language at any level or can speak it in any real situation. Even this is a great advance on tests made by supporters of the first position because all their tests indicate is whether a student can do the test or not. There is no implication as to his general performance which is what the test is supposed to be an indication of. It is logically possible to learn a good deal about Latin grammar and prepare a number of average proses and yet not cope with a prose test. The reason for this is that the relation between one prose and another is not clear, that unless great pains are taken to limit the vocabulary and control the grammar such a test is just another exercise with no content and little predictive validity.2

Those who take up the third position we have spoken of hold a united view of language and language learning. They see language as creative and at the same time rule-based (Chomsky, 1965). Language

¹ See below, Chapter 2.

² See below, pp. 8-9.

and learning are inter-dependent because no one ever stops learning language, especially his own, because there is no such thing in the world as a state of language (cf. de Saussure, 1915) separate from people and therefore from learning. Attractive as this view of language learning is there are difficulties connected with it. No one ever really said, no linguist, traditional or structuralist, that language in the world is separate from people and learning. The language they describe is an abstraction and the language the transformational grammarians describe is also an abstraction which is as much divorced from learning in its formalization as any previous view or model.

It is really too early yet to see the influence those who take up this third position will have on teaching. However, three assumptions are already being made. The first of these is, as Carroll points out in Chapter 4, that it is competence that is being aimed at. Now since underlying all surface realizations is Deep Structure (cf. Chomsky, op. cit.) then what may well be important for the student to do is to learn the Deep Structure and the transformations that may be employed on it. The second assumption is that the L2 and LF learner will need to make the same kind of approach as the L1 learner. Now we still don't know which approach the L1 learner makes (cf. Lyons and Wales, 1966) but since the transformational generative approach is directed towards the way a native speaker learns his own language, its formalization will be in L1 terms and any teaching making use of it must also approach the L2 and LF learners in the same terms. The third assumption is linked to L1 learning too since it accepts the demands of situation, i.e. the environment in which language is used. Now this is not a necessary part of the theory but since its emphasis is on L1 language learning the situation must be taken into account in preparing teaching materials.

Here there is a close link with yet another view of language, that of the Firthians (cf. Robins, 1964). J. R. Firth's own approach to language description is probably to be reckoned as structuralist but he was greatly concerned with extra-linguistic features, with situation and its link with language which, after Malinowski, he termed context of situation. Now it may be, as Chomsky has argued, that situation cannot be categorized (Chomsky, 1966). But, as Firth maintained, in any community there is to be found a whole range of different languages, a network of sub-languages which interlock at different points so that any one speaker will control only some of these sub-languages or varieties and will certainly make use of different ones for different purposes and on different occasions. While the linguist may restrict his theoretical concerns to abstractions, to well-formed sentences and the like, the learner and the teacher will be as much concerned with the learning and with the situation. The linguist's real concern is

with the state of the language, an abstraction divorced as much from the learning as from the situation. Firth's emphasis on situation, therefore, is no less important or relevant than Chomsky's on learning. While both may be difficult for the linguist to formalize, they provide valuable hints and starting points for the teacher and tester which they may follow up in their own work.

The influence on teaching of the view of language held by those who take up this united position is likely to emphasize tasks and problems rather than patterns and repetitions, strategy rather than memory. The argument runs thus: the L1 learner learns his language in some 'principled' cognitive way, not by storing every possible sentence type in his head, so the L2 learner needs to be taught in some way that simulates this procedure. And, so the argument continues, the L1 learner learns, again not by being fed lists but by meeting problems. making mistakes. So the L2 learner must be provided with problems for solving. The aim of the teaching is to provide least assistance, so to structure the problem that good learners need no assistance. Teaching and testing meet here since the measure of achievement depends on how much assistance is demanded (cf. Williams, 1961). And the argument concludes, since the L1 learner always learns in a situation so the L2 learner must be provided with meaningful, contextual exercises. L2 teaching attempts to simulate in a structured way the L1 environment. The reader will have noted the links implied in this paragraph between teaching and testing and programming. The potential strength of PL lies in its almost infinite use as a method of arranging and trying out hypotheses as to learning structures.

It should be clear by now that the good test is an obedient servant since it follows and apes the teaching; this is particularly true for the Achievement test (see below p. 7). There will be a great difference between our second and third positions in so far as they influence teaching. The difference as regards testing will not be so obvious. A third position test will be much more situational because the teaching is. It will be less concerned with 'pure' language features, e.g. segmentals, stress, because the teaching is less so concerned; it will be as much concerned with objectivity because it is also a good test. What it will do that is different from a second position test is to set 'problems' which are not all that different from the translation problems of the first position test but are directly related to the corpus which has been taught. It will set tasks such as comprehension which need resolution actually during the task and are not channelled through the stimulusresponse funnel of question and answer after the passage. What is more it will probably attempt to tap the L2 learner's creativity which our third position teaching will have re-directed by allowing the learner to make use of the rules he has learnt in order to establish how far

he has internalized them. This is very different from a multiple choice answer to a stimulus involving, say, intonation, because what is being demanded is use of the rules to produce a sentence or 'response' which is new. There is no reason why such a creative new sentence should not be incorporated within an objective multiple choice framework, thus:

Example 1 (teaching of the Imperative in English has been completed)
Jane and her brother, Tom, are waiting for the bus home after school. 'Let's go swimming,' says Jane. But Tom has another idea. What does he say?

- 1. Swimming?
- 2. Don't let's go swimming, let's . . .
- 3. Let's go swimming.
- 4. You go swimming.
- 5. Don't go swimming, go . . .

It may be objected that it is some kind of translation that is being tested here or, if not translation, that learners are being expected to think in terms of formal categories. This may be so. What is being claimed is that as well as relating test to rule, it also makes situation important. Such a test could well be drawn up in picture format which would cut out all need for explanation.

Compare this with a second position test.

Example 2

Tom does not want to go swimming with Jane but he does want to go somewhere else with her. What does he say to her?

- 1. Do let's not go swimming.
- 2. Let's go swimming.
- 3. Don't let's go swimming.
- 4. Don't go swimming.
- 5. You go swimming.

2. LANGUAGE TEST USES

There is some confusion in the use of terminology in Language Testing and it may be useful to discuss some of the conflicting terms. They occur especially in the area of test uses, so that the terms achievement, attainment, proficiency, diagnosis and aptitude are often confused. There seem to be three main usages, to two of which the readers of this volume are exposed. The first usage links Proficiency and Aptitude, Achievement and Attainment, and has, as all three usages have, Diagnosis separate. Thus we speak of Proficiency (Aptitude) for or in something to do something else; we speak of Achievement (Attainment) in something by itself; and we speak of Diagnosis of something. Thus in this usage Proficiency (Aptitude) tests the student's present ability for future learning. Achievement (Attainment) tests his present knowledge as

indicative of past learning, and Diagnosis is the teacher's concern for what has gone wrong.

The second usage is the present writer's. It distinguishes four uses, combining Achievement and Attainment, in terms of time and of subject matter. Thus if we use an arrow of time going both ways and if X is the test, then we may symbolize as follows:

Achievement or Attainment tests are concerned with assessing what has been learnt of a known syllabus. This may be within a school or within a total educational system. The use of Achievement tests is to find out how much has been learnt. They are concerned, therefore, entirely with the past though they often are used for predictive purposes. That they should be used predictively, i.e. to make claims about future performance, is probably inevitable though it is not the function of an Achievement test to predict. All an Achievement test can do is to indicate how much of a syllabus of what has been taught has been learnt. It cannot make predictions of itself as to pupils' future performance unless the syllabus has been deliberately designed, as it should be, with the demands of teaching and learning in mind.

If Achievement tests are concerned entirely with the past, which explains the arrow only going one way, then Proficiency tests are concerned both with the past and with the future. But there is no syllabus control over the past, as with Achievement tests, and so the test constructor cannot draw on the syllabus for the content of his test. Instead he must make up his own syllabus. A good Proficiency test is, or implies, such a syllabus. This lack of a syllabus is indicated by the broken arrow. The Proficiency test also looks forward, i.e. it is used to predict skill in language for some extra-linguistic purpose, e.g. proficiency in English for something else. Hence the symbol Y.

An Aptitude test also looks forward, but to a language skill. Thus it is concerned with assessing skill in e.g. language for language. This assumes that skill for future language learning, proficiency, say, in acquiring a second or third language, should and can be measured in existing first language skill. The uncertainty here is indicated by the brackets round the first X at the side of Aptitude. Notice that Aptitude, unlike Proficiency, has no interest in the past.

Diagnosis does not fit readily into this scheme of uses because it relates entirely to the use made of the information and not at all to the presence of a skill in the learner. That is to say we can speak of a learner's Achievement, Proficiency and Aptitude but not, of course,

of his Diagnosis. A Diagnostic test is for teacher consumption; it is a use made by a teacher of the information provided by one of these skills, Achievement, Proficiency and Aptitude. A Diagnostic test may be constructed for itself or it may be an additional use made of an Achievement or Proficiency test. If it is specially constructed it could be argued that the learner's absence of skill is being tested so that another name for a Diagnostic test might be a Non-Achievement test. There is the further argument that such a specially constructed test is concerned with establishing systematic errors and not at all with randomly occurring mistakes so that it is not absence of skill but difference of skill between L2 learners at different stages of learning which is being tested.

This discussion of terminology deals with uses of tests. Under use we have considered Achievement, Attainment, Proficiency, Aptitude and Diagnosis. Kinds of test would include such terms as Oral tests, Writing tests, Comprehension tests, L1 tests. All the uses considered here could be made of one kind of test. An oral test could be used for achievement, diagnostic or even aptitude purposes. It all depends on what the user of the test, i.e. the teacher actually making the particular use, is looking for.

The third usage is that of Ingram (in Chapter 5). What she does is to use Attainment as a major class of which Achievement and Proficiency are both members. Since, as we have seen, Achievement and Proficiency uses are closely related and since there is the special Proficiency case of Pre-Attainment in which a test samples the content of a course which students wish to enter (cf. Mialaret and Malandain, 1962; Davies, 1965a, p. 48), it may well be that such a classification is the most useful. Ingram uses Aptitude and Diagnosis as described above.

3. EVALUATION IN LANGUAGE TESTING

As well as concerning itself with language and with learning language testing is also very much concerned with evaluation. As part of Psycholinguistics it is experimental and therefore must make use of the appropriate statistical procedures. What these do (see Chapter 2) is to ensure the Reliability and Validity of the test. Full discussion of these ideas is given by Pilliner in Chapter 2. But it is worth emphasizing here the equal partnership in language testing of evaluation with linguistics and psychology.

Pilliner (Chapter 2) deals with the meaning of validity and its implications in evaluation. He discusses three kinds of validity: Predictive, Concurrent and Content. A fourth kind, Construct Validity, is sometimes invoked. The distinction he makes (p. 31), following

Wiseman, between the syllabus-content approach and the goal-oriented approach to a test is our distinction between Achievement and Proficiency tests. The Achievement test is syllabus-content based; the Proficiency test is related to the demands and aims of the learning process. This relatedness can be measured quantitatively, as Pilliner describes, drawing on Predictive or Concurrent Validity. It can also be assessed logically by Construct Validity. Where Pilliner uses Content Validity we would use both Content and Construct. In addition there is the fifth kind, Face Validity. This may appear trivial since it refers to acceptance by the layman. But in education it is often important to show as well as know that what is being done is relevant. Face Validity may therefore not be so trivial after all.

We now can produce the following scheme for validities:

Type	related to	Criterion	
Face		Lay view of language	
Content		Syllabus analysis	
Construct		Theory of aims in Language Teaching	
Predictive		Later test of language ability	
Concurrent		Another test or measure (e.g. teacher's assessment) of language ability given at same time	

TABLE I

Ingram points out (Chapter 5) that it is only after looking at the statistics that we know the worth of a test. However, there are degrees of knowingness and they differ for the different uses and for the different validities. There is a relationship of likelihood between test use and validity so that it is likely that for an Aptitude use the test constructor will be most concerned with a theory of language learning and therefore with Construct Validity; and at the other end of the scale, for a Proficiency use he will be most concerned with Concurrent or Predictive Validity. Now Predictive Validity is the most demanding statistically of all the validities since it seeks to establish the adequacy of a test by relating its scores to a criterion of future performance, the real 'pay-off' of any test. Concurrent Validity is also established statistically

but does not expose the test's scores to such hard probing, confining its correlations to parallel or already existing tests. None of the other validities demands statistical procedures though they may use them. Hence the Scale below (Table 2) where we represent our over-simplified connexions between test uses and kinds of validity and suggest a linking scale of Statistical Revelation. What this implies is that there is a further relationship of likelihood between test use and type of validity such that some test uses demand more statistics than others.

Use	Validity		
Aptitude	Construct	Less	
Achievement	Content	ľ	Statistical
Diagnosis	Content		Revelation
Proficiency	Concurrent		
Proficiency	Predictive	↓	
		More	

TABLE 2

All test uses will still need the run of the mill statistical analysis for item analysis, reliability and inter-test correlation. Measures of central tendency, means and standard deviations are necessary for comparative purposes and for assessing the comprehensiveness of the test. Even so, it is, as Pilliner and Ingram point out, validity that really matters. So much does it matter that it is possible to imagine a test with very low item discrimination because many heterogeneous abilities are being tested, poor reliability because the range is small; and yet such a test could have high validity, and this is what counts. It is the test constructor's assumptions in language learning that are really being analysed. A good test is a device for framing these assumptions and as much a learning device as a language laboratory or a programmed text.

4. LANGUAGE TEST ANALYSIS

Various kinds of Language Test analysis have been suggested. They all depend on the analyst's view of language. Both Carroll and Ingram (Chapters 4 and 5) provide schemes which involve such analyses. One scheme they do not discuss follows on from a discussion of language-performance by Corder (1966). This links a Piagetian developmental scheme with a language level plan. The balance is provided by the *channels* of production and reception though many