

New Directions in Language Testing

Edited by

**Y P Lee, Angela C Y Y Fok, Robert Lord
and Graham Low**

New Directions in Language Testing

*Papers presented at the International Symposium on
Language Testing, Hong Kong*

Edited by

***Y. P. Lee, Angela C. Y. Y. Fok,
Robert Lord and Graham Low***

University of Hong Kong



PERGAMON PRESS

Oxford · New York · Toronto · Sydney · Paris · Frankfurt

U.K.	Pergamon Press Ltd., Headington Hill Hall, Oxford OX3 0BW, England
U.S.A.	Pergamon Press Inc., Maxwell House, Fairview Park, Elmsford, New York 10523, U.S.A.
CANADA	Pergamon Press Canada Ltd., Suite 104, 150 Consumers Road, Willowdale, Ontario M2J 1P9, Canada
AUSTRALIA	Pergamon Press (Aust.) Pty. Ltd., P.O. Box 544, Potts Point, N.S.W. 2011, Australia
FRANCE	Pergamon Press SARL, 24 rue des Ecoles, 75240 Paris, Cedex 05, France
FEDERAL REPUBLIC OF GERMANY	Pergamon Press GmbH, Hammerweg 6, D-6242 Kronberg-Taunus, Federal Republic of Germany

Copyright © 1985 Pergamon Press Ltd.

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means: electronic, electrostatic, magnetic tape, mechanical, photocopying, recording or otherwise, without permission in writing from the publishers.

First edition 1985

Library of Congress Cataloging in Publication Data

International Symposium on Language Testing (1983
Hong Kong)

New directions in language testing.

(Language teaching methodology series)

1. Language and languages—Ability testing—

Congresses. I. Lee, Y. P. II. University of Hong
Kong. III. Title.

P53.4.I58 1984 428'.0076 84-1082

British Library Cataloguing in Publication Data

International Symposium on Language Testing
(1982: Hong Kong)

New directions in language testing—

(Language teaching methodology series)

1. Language and languages—Ability testing.

I. Title II. Lee, Y. P. III. Series

418'.007 P53.4

ISBN 0-08-031535-6

Other Titles of Interest

ALDERSON, *Evaluation* (Lancaster Practical Papers in English Language Education, vol. 6)

CARROLL, *Testing Communicative Performance*

See also: **SYSTEM, The International Journal of Applied Linguistics and Educational Technology**

ESP Journal, An International Journal of English for Specific Purposes

Free sample copies available on request

Acknowledgements

The Editors wish to thank the British Council for providing sponsorship for two of the keynote speakers and for undertaking some of the coordinating work on behalf of the Organizing Committee of the International Symposium on Language Testing. They would also like to thank those who willingly and generously accepted key roles: Kenneth Topley, former Secretary for Education in Hong Kong, who opened the Symposium; Alan Brimer, Dean of the Faculty of Education, University of Hong Kong; Francis Johnson, Chairman of the Department of English, Chinese University of Hong Kong, and Oliver Siddle, Representative of the British Council in Hong Kong, who chaired the Plenary Sessions; as well as the many others who chaired the working sessions: Peter Barnes, Nelson A. Berkoff, James D. Brown, Evelyn Cheung, William Cheung, Peter Falvey, Anne Hilferty, John Hunt, Keith Johnson, Randall Jones, Rex King, Helen Kwok, Teresa Leung, Alan Maley, Elaine Marshall, Michael Milanovic, Alan Moller, Pauline Rea, John Read, Douglas Stevenson, Merrill Swain, Tsao Feng Fu, Christine Wright and Arne Zettersten. Nor would they want to overlook the great deal of painstaking labour that went on behind the scenes in making the International Symposium a success: Amy Lam for her work as Treasurer to the Organizing Committee; Christine Wright, Evelyn Cheung, Cynthia So, Caroline Lo, Helen Yeung, and Irina Lau, who took care of the details of the programme and registration; Matthew Leung, Amy Lam, Cheng Ngai-lung, and Ip Kung Sau, who provided information and planned the amenities; Tsao Feng Fu, who was the Translation Adviser; and Joe Sing He and Edward Leung, who provided technical and audio-visual assistance. Our especial thanks, finally, to Connie Kwok who typed the final draft.

List of Contributors

NELSON A. BERKOFF, The Hebrew University, Jerusalem.

ALAN DAVIES, University of Edinburgh.

ANGELA C. Y. Y. FOK, University of Hong Kong.

PATRICK E. GRIFFIN, University of Hong Kong.

GUI SHI CHUN, Institute of Foreign Languages, Guangzhou.

RANDALL L. JONES, Brigham Young University, Utah.

NANCY LEE, University of Hong Kong.

Y. P. LEE, University of Hong Kong.

GRAHAM D. LOW, University of Hong Kong.

DONALD M. MORRISON, University of Hong Kong.

THOMAS M. PENDERGAST, JR. Osaka University of Foreign Studies.

PAULINE M. REA, University of Dar Es Salaam.

DOUGLAS K. STEVENSON, Universität Essen.

MERRILL SWAIN, Ontario Institute for Studies in Education, Toronto.

TIBOR VON ELEK, University of Göteborg.

YANG HUI ZHONG, Jiaotong University, Shanghai.

ARNE ZETTERSTEN, University of Copenhagen.

Editors' Introduction

In December 1982 Hong Kong hosted an International Symposium on Language Testing. Its organizers, drawn from the Language Centre of the University of Hong Kong, chose two themes which have been, in microcosm, of central concern to teachers and testers in the Language Centre and, in macrocosm, to language testers the world over. The two themes chosen were: *direct or performance* testing, and *large-scale* testing. Hong Kong offers a common focus for these two areas of language testing, which otherwise could seem disparate and none too obviously related.

In recent years there has of course been a general shift in emphasis in approaches to second or foreign language teaching, with a language being increasingly perceived as a tool to be used in a purposeful and controlled manner. With this shift have emerged new ideas about language testing and new ways of evaluating the performance of second language learners. A concern with measuring how well language is used within certain contexts and as a means of establishing aspects of context has therefore become one of the main features of this new direction. To date, a good deal of work has been devoted to identifying and formalizing the required target behaviour of learners at the end of their courses, and to constructing viable formats with which to test what one might call "proficiency in action". Until recently there has been relatively little published work examining the data that results from such tests; for example, the construct and predictive validity of such data, and ways of maintaining reliability.

Hong Kong shares with many other multilingual societies the problem that the medium of instruction is not the first language of the vast majority of its school population. As a direct result of this, second language achievement, and the evaluation of second-language achievement play a crucial role in educational planning. Language attainment needs to be carefully assessed at every stage, and students' ability to use the second language for learning carefully monitored. Without a detailed knowledge of the students' ability to make use of the medium of instruction, no appropriate plans for improvement and development can be formulated. It is for this reason that, in exploring new directions in the design of language tests, the validity and reliability of the testing instruments need to be established with particular care. On the other hand, such procedures can be time-consuming and this has to be balanced against practicability, which tends to mean simplicity of test administration and ease of scoring, especially when the tests are administered to very large groups.

It was our hope that the shared experience and aims of participants in the Symposium would lead to new ways of producing suitable tests to evaluate the language behaviour of the student population as well as shed new light on some of the perennial problems besetting our various educational systems.

The participants came from 16 different countries. The Symposium consisted of papers, workshops and demonstrations, and included four plenary sessions with invited speakers: Alan Davies (University of Edinburgh), Randall Jones (Brigham Young University, Utah), Pauline Rea (University of Dar Es Salaam), and Merrill Swain (Modern Language Centre, Ontario Institute for Studies in Education, Toronto).

The Opening Ceremony included addresses by the Hon. K. W. J. Topley, then

Secretary for Education in Hong Kong, and Robert Lord, Director of the Language Centre and Chairman of the Symposium Organizing Committee.

This book consists of an edited selection of 15 of the papers presented, and includes all four of the papers presented at the Plenary Sessions. The papers have been reorganized to form four sections, which differ slightly from the three sections of the Symposium. Since virtually every paper is concerned with the question of performance testing, it seemed more sensible to remove this as a separate section. On the other hand, so many papers were concerned with questions of test validity, that it was decided to group these together under the general heading of *Validating Language Tests*. This section is organized from the more general to the more mathematical and the last paper in particular, that by Griffin, is clearly mathematical in orientation. It contains recent work on virtually every major dimension of validity: predictive validity (Low and Lee, Fok), face validity (Stevenson), and construct validity (Lee). Papers on content validation (e.g. Morrison and Lee) have been included in other sections. Similarly, von Elek's discussion of the possibility of validating self-assessment tests has been placed in the section on *Large-scale Testing*.

The section devoted to *Large-scale Testing* has been retained and contains attempts to solve the problem of how to reconcile recent ideas on communicative testing with the reality of large numbers of students and the little time and or money available for testing each individual student. Swain considers the case of classroom teachers wishing to evaluate the effectiveness of their bilingual education programmes (here French and English) in Canada and the development of a package involving simulated tasks which is administered by classroom teachers. Von Elek discusses recent work on the long-running project to develop self-assessment tests for adult learners of English in Sweden, particularly adult migrants. Zettersten discusses work done in Denmark to develop large-scale tests of lexical skills, with a special focus on ways in which such tests can be computerized. Yang and Gui, in a similar spirit, discuss recent work on the development of a national language proficiency test (of English) for young adults in China.

The section *Testing Oral Skills* begins with an overview of the situation by Jones, who cites numerous examples of recent, particularly American, work in the area. This is followed by three case studies. Both Lee and Morrison, and Berkoff give details of attempts to construct and validate tests aimed at measuring the oral skills of university students by means of a group testing approach. The paper by Pendergast is of a rather different nature and has closer links in many ways with the large-scale testing section; it is a summary of work done over the last few years in Japan in developing a very rapid (2–3 minute) test of oral ability where the examiner is aided by a "testside" computer.

We have taken the two plenary sessions which were most general in scope and put them at the front of the book to form a readable though detailed introduction to the current language testing scene. Both are concerned that testing should not be seen as something divorced from language teaching. Davies focuses particularly on a number of recent errors of philosophy or judgement, while Rea is concerned to develop a method of handling test types which is usable by the classroom teacher.

We hope, therefore, that the papers published here will be of interest and use to a wide variety of readers, from research workers attempting to develop new ways of designing and validating tests, to language teachers who would like to know what has been going on over the last few years and who would like a framework for dealing with the multiple variations on the theme of communicative or direct or performance testing which have been appearing in the literature.

Contents

List of Contributors	ix
Editors' Introduction	xi
1. Language Testing and the Curriculum	1
Follow my leader: Is That What Language Tests Do? ALAN DAVIES	3
Language Testing and the Communicative Language Teaching Curriculum PAULINE M. REA	15
2. Large-Scale Testing	33
Large-scale Communicative Language Testing: A Case Study MERRILL SWAIN	35
A Test of Swedish as a Second Language: An Experiment in Self-assessment TIBOR VON ELEK	47
The English Proficiency Test Used in China: A Report YANG HUI ZHONG AND GUI SHI CHUN	59
Experiments in Large-scale Vocabulary Testing ARNE ZETTERSTEN	67
3. Testing Oral Proficiency	75
Some Basic Considerations in Testing Oral Proficiency RANDALL L. JONES	77
Simulating an Academic Tutorial: A Test Validation Study DONALD M. MORRISON AND NANCY LEE	85
Testing Oral Proficiency: A New Approach NELSON A. BERKOFF	93
OLAF N. 73: A Computerized Oral Language Analyser and Feedback System THOMAS M. PENDERGAST, JR	101

4. Validating Language Tests	109
Pop Validity and Performance Testing DOUGLAS K. STEVENSON	111
How Shall a Test be Referenced? GRAHAM D. LOW AND Y. P. LEE	119
Language Proficiency and Related Factors ANGELA C. Y. Y. FOK	127
Investigating the Validity of the Cloze Score Y. P. LEE	137
The Use of Latent Trait Models in the Calibration of Tests of Spoken Language in Large-scale Selection-placement Programs PATRICK E. GRIFFIN	149
Bibliography	163

Language Testing and the Curriculum

Follow my leader: Is That What Language Tests Do?

ALAN DAVIES

An Agony in Eight Fits*

Fit the First

*"Just the place for a Snark! I have said it twice:
That alone should encourage the crew.
Just the place for a Snark! I have said it thrice:
What I tell you three times is true"*

In disciplines where data reveal themselves to consistent view and leave no case to appeal, where objectivity, what Melville called the level log and line, holds sway, in scientific disciplines such as medicine, what students learn is, we recognize, true. It may be that in different medical schools there are varying emphases such that a young surgeon from one medical school may be more practised with the knife while one from a second school may be more into computers and data banks. But we are confident that they share a common core such that the diagnosis of an ailment will be identical whichever surgeon makes it.

And yet how true is this? Our emotional selves may want to believe that medical opinion is not only consistent but true as well, but our rational selves, and our experience when we listen to it, reminds us that the commonsense world is a hit-and-miss affair, even in rubbery hard sciences like medicine. What doctors do, in fact, in order to overcome the subjectivity and unreliability of their diagnoses is to establish very strong consensus and to ensure that their apprentice youth are firmly socialized into that consensus. Thus, it does not matter in medicine to be wrong, what matters is to be odd. And since being wrong has probably no useful meaning, since we do not know what being right would be, even if we take a realist position, patients, potentially all of us, are no doubt grateful that consensus validity can be achieved, promising at least some small kind of content or concurrent validity.

Paul Atkinson (1981) describes the socializing into consensus of young doctors in the Edinburgh Medical School. Atkinson discusses his fieldwork as part of a critique of young medical students during the clinical phase of their training. As he points out, "clinical" phases of training rest on the assumption that the trainee is to gain practical knowledge and experience through some form of exposure to, immersion in, and some sort of "practice" on "real" settings (p. 115). What he tries to show is how this clinical reality is socially organized, achieved and managed.

"... The medical reality to which the beginning student is exposed is by no means straightforward. There is nothing natural about such 'reality'. Like any other it is a matter of

* *Agony* is glossed by Carroll as a struggle that involves great anguish, bodily pain or death. It is possible that Carroll also had in mind the "woeful agony" that periodically seizes Coleridge's *Ancient Mariner*, forcing him to tell to strangers his "ghastly tale".

social construction. What novitiate students are introduced to, then, is not the practical reality of some essential clinical medicine. Rather, they encounter the dramatic enactment of a particular form or version of medical work. . . . I found myself identifying, at least tentatively, a style of teaching I labelled 'thaumaturgical' (wonder working). The patients and medical students would find themselves recruited partly as audience, partly as stooges, partly as props in displays of clinical expertise and diagnostic skill. The effect of such displays, and of less dramatic teaching too, is to produce vivid firsthand reconstructions of the logic and rationality of clinical medicine. It is as audience to these performances that the students gain their first clinical experiences."

In clinical teaching what we have, says Atkinson, is a mock-up in which the "facts of the case" are determined and legitimized by reference to the procedural rules of "correct" and "methodical" enquiry. And he goes on to suggest that this is true of all science, reminding us of the structure of scientific thinking T. S. Kuhn has described, in which what is at any time "true" is only so because, or if, it fits into the currently accepted paradigms. No doubt Kuhn, and his followers, like Atkinson perhaps, would as empiricists have been delighted to read the 1901 parody version of *MIND* (published as *MIND!*) in which the pragmatist philosopher F. C. S. Schiller provided a mock-serious commentary on Lewis Carroll's *The Hunting of the Snark* which, Schiller insisted, is a satire on the Hegelian philosopher's search for the Absolute. It is from *The Hunting of the Snark* that I take my texts for this talk. No doubt, in philosophy, empiricism is a comforting refuge, for there no patients die, no bridges fall down, and no languages remain unlearned. But in the world of applied (social) sciences in which we deal we are genuinely concerned that what we do should somehow be in pursuit of some external validity represented by consensus, and that validity should not be solely a matter of internal and therefore self-justifying consistency.

Fit the Second

*"Come, listen, my men, while I tell you again
The five unmistakable marks
By which you may know, wheresoever you go,
The warranted genuine Snarks"*

If we accept that reality is thus socially constructed even in hard sciences, how much easier it is to agree that in the softer social sciences and, in the quicksands of Applied Linguistics and Language Teaching, facts (like snarks) keep changing shape.

"Reality" for most language learners is socially constructed. However reconstructed the syllabus, the teaching and anything else may be, what the students' gaze (and the public's) is fixed on is the test, no matter how unreconstructed that may be. And, understandably, learners are grateful for the social rather than individual reassurance, because what they want is the reassurance that can only come from consensus: thus it would be possible to contemplate a complete mismatch in which the individual succeeds communicatively, interactively, but is dissatisfied because he does badly on the test which is out of phase with "true" measures. Eventually, such a mismatch represents a lack of agreement as to what stands for the language. Examples of such a mismatch are found for example in situations where the syllabus and the teaching may be wholly concerned with the language, may be communicative, but where the test is still dedicated to literature. A student who is unprepared for such a literary test (which is institutionalized in the educational system) will not be grateful for a syllabus and a teaching method that make it possible for him to communicate with ease. He has not been able to unsocialize his view of what language learning goals in his situation are – they are literary and have to do with

some kinds of language-like behaviour rather than with language behaviour.

I want to go on to some specific examples of such mismatches and the harm they do.

Fit the Third

“‘That’s exactly the method’, the Bellman bold
In a hasty parenthesis cried,
‘That’s exactly the way I have always been told
That the capture of Snarks should be tried!’”

A test situation such as the one just described is not satisfactory. Such situations fall into two main problem categories: problems of excessive conservatism or problems of unthinking radicalism.

First, two conservative examples. One of these is a historical example which illustrates how a language test can become ritualized and irrelevant except in a symbolic way. The example is that of the *neck verse* which is defined in the Shorter Oxford Dictionary as

“Neck Verse 1450. A Latin verse printed in black-letter (usually the beginning of Psalm 51) formerly set before one claiming benefit of clergy by reading which he might save his neck. Now only Historical” (sic!)

(Psalm 51 – for those who wish to put something aside in the hope that they can, if it ever proves necessary, preserve their necks – begins “Miserere mei, deus”, or in our present vulgar or doric “Have mercy upon me, O God, after thy great goodness: according to the multitude of thy mercies, do away mine offences.”) The neck verse may be “only *Hist.*”, but we do still make judgements on similarly ritualistic foundations, e.g. the language question in the UK Census presented in Wales and Scotland: “Do you speak Welsh/Scots Gaelic?” (only recently added to by the secondary questions: “If so, do you read/write it?”). Or, for example, this item from a questionnaire carried out in Bedford and reported in Sutcliffe 1982 (*British Black English*). Sutcliffe was interested in the variability of English among the younger black community in Bedford. His questionnaire attempted to find out what percentage of the sample was able to speak Broad Creole.

“The test sentence used was the Jamaican Creole: *mi asks di man fi put mi moni iina him pakit* (I asked the man to put my money in his pocket). Subjects were asked to translate it and were asked: ‘Do you sometimes speak like that?’ A surprising number . . . said they did – 78 % . . . Similarly, over 90 % . . . reported using at least some Creole” (p. 151).

It is not surprising that Sutcliffe’s results have been challenged – as he himself admits. It is difficult to see *what* conclusions could be drawn from so unsatisfactory a test item.

My second conservative example comes from Saudi Arabia (personal information, Alan Beretta, 1982) in the British Aerospace English Language programme, where what is required is to obtain a “pass” in the English Comprehension Level (ECL) examination. The only course materials used are the American Defence Language Institute’s ALC (American Language Course). Over the years there have been complaints about and criticisms of this course but, I am assured, while it might have been possible to modify it in practice, the rigidity of the ECL examination does not allow for any change. Not only do the authorities forbid any change in the examination, but the students themselves know exactly what is demanded, and insist that no changes in the course be made.

Approximately 70 % of all items test the students' capacity to associate tenuously related lexical items, e.g.

Stem: A pioneer was a tough man.

Choices: tradesman, colonist, cowboy, foreigner

where the correct answer is *colonist*. The effect of such rigid testing is that *pioneer*, *settler*, *colonist*, *explorer* are taught as synonymous (known as *same-same* to the students), and that much of the teaching consists of similar vocabulary lists. My informant concludes that the format of the test means that there is no teaching of writing, speaking, reading of more than sentence-length, or listening to anything of more than sentence-length. He provides this science-fiction scenario of a "typical" teacher-student interchange:

Teacher: pick up the book

Student: pick up same-same collect same-same take away

T. start the next sentence

S. start same-same begin same-same take up same-same commence

T. OK that's enough of that

S. enough same-same adequate same-same sufficient

T. keep quiet

S. keep quiet same-same shut up same-same quit it, same-same can it, buddy . . .

The neck verse test and the same-same test both have residual validity. Those who were both poor and illiterate would not have been priests and would, in the absence of any kind memorizer, have failed the test. Similarly, the same-same test keeps close to the ALC syllabus. What disturbs us about both is precisely the conservatism that characterizes them. They *ought* to have been changed, to relate to more valid ideas of literacy, to help revise the ALC course in such a way as to prevent what, my informant tells me, happens every course end where students cheerfully proceed to the next phase of training, linguistically incompetent in English after a year (or often 2 or 3 years) of intensive study. No doubt it is a virtue of achievement tests to stick close to the syllabus – as has often been said, their job is to sample, and the validity issue is one for the syllabus. But without change in the syllabus that virtue of the achievement test is a recipe for disaster: the teachers alone just cannot generate enough power to bend those institutional bonds.

Fit the Fourth

"For the Snark's a peculiar creature that won't

Be caught in a common-place way,

Do all that you know, and try all that you don't:

Not a chance must be wasted today."

Let me now take my two unthinking radical examples. The first comes from Malaysia where some years ago the Curriculum Development Centre (CDC), concerned to integrate the two language-medium systems, English and Malay, as quickly as possible, introduced a new Communication Syllabus for English into the secondary schools. The idea was that what secondary school graduates needed was the ability to communicate in English, particularly in spoken English, and of course by communicate was meant that

rather special *communicative* view which involves the capacity to cope adequately with the English of every possible situation. I think it is now recognized that such a view is one kind of vague description of what the native speaker can do. Schools were provided with the new syllabus which was mandatory: it consisted largely of sociologically described situations, e.g. in a tourist office or hotel, with no indication of what language was relevant or necessary. There were no textbooks. The Examinations Syndicate, a separate organization from the CDC, were unwilling to radicalize their school certificate type examination to bring it in line with the new syllabus. The result was a grotesque mismatch in which the syllabus was in large part unteachable because it properly required, in part, the competence and the intuition of a native speaker teacher and the examination did not in any case test the syllabus. Failure was severe except among those from the more élite English-medium schools; many average English-medium students and most of the Malay-medium students failed. Remember that the purpose of the new syllabus was to help those very Malay-medium students who had had no access to English medium, and that English medium was then on its way out. Twenty-five per cent of the new syllabus was intended for spoken English, again to encourage the use of spoken English among those with least opportunity in school to use it. The Exams Syndicate refused to give more than a token share of the examination percentage to oral English. The CDC had acted with brave but foolhardy zeal: it was right to be concerned about promoting appropriate use of English, but wrong not to check with its sister, but separate, examinations institution, and with the teaching profession. Change is essential but it needs a fabian lead.

The same comment can be made on my second unthinking radical example, this time from West Africa. In the 1960s, D. W. Grieve wrote an excellent report on English examining in the West African secondary schools. He was then asked to rewrite the examinations, and at that point things began to go wrong. Grieve's interest in linguistics, an interest which had helpfully informed his critique of the existing school certificate (modelled on the Cambridge overseas school certificate), dominated the new examination which he constructed. Whether or not the new examination was piloted I don't know – if it was, it can only have been in the main urban centres, a warning to us all when we are sampling whole educational systems. But although syllabus and text books were redrafted and revised, the change was too great and the teachers could not cope. Of course all systems have inertia and some way round the problem of mismatching is typically found, if only, as in Singapore, by insisting with the weight of authority, on reverting to a traditional school certificate type of examination so as to encourage (by backwash) in the schools continuous writing in English which, so it seems, was being dropped after the introduction into the examination of a more objective-type approach.

Change in language teaching must be possible; that is, there must be some way of responding to new ideas and demands. It is best if the change comes in through the syllabus and the examination and the teacher. If a choice has to be made among these in order to move quickly, then undoubtedly the test/examination is the most sensitive; it is the most controllable, it acts overall, it is most difficult (*pace* W. Africa) to ignore, it has most certainty in terms of its goals. The test/examination is a major and a creative influence for change and development in language teaching, and if there is a need to choose, then that is what should always change first. *But* the influence for good that the test/examination can have (and I would cite here the work of the Royal Society of Arts of London on its language teachers' examinations, its communicative test of EFL, its new examination in ESL; the Joint Matriculation Board Test in English (Overseas); the