Fundamental Considerations in Language Testing

语言测试要略

Lyle F. Bachman

上海外语教育出版社

牛津应用语言学丛书

语言测试要略

Fundamental Considerations in Language Testing

Lyle F. Bachman 著

上海外语教育出版社

上海市版权局
著作权合同登记章
图字:09－1999－035 号

牛津应用语言学丛书

Fundamental Considerations in Language Testing

语言测试要略

Lyle F. Bachman 著

上海外语教育出版社出版发行
（上海外国语大学内）
深圳中华商务联合印刷有限公司印刷
新华书店上海发行所经销

开本 880×1187 1/32 13.25 印张 513 千字
1999 年 4 月第 1 版 1999 年12月第 3 次印刷
印数:1500 册

ISBN 7－81046－573－2
H·584 定价:26.00 元

Oxford University Press
Great Clarendon Street, Oxford OX2 6DP

Oxford New York
Athens Auckland Bangkok Bogota Bombay
Buenos Aires Calcutta Cape Town Dar es Salaam Delhi
Florence Hong Kong Istanbul Karachi Kuala Lumpur
Madras Madrid Melbourne Mexico City Nairobi
Paris Singapore Taipei Tokyo Toronto

and associated companies in
Berlin Ibadan

Oxford and *Oxford English* are trade marks of Oxford University Press

ISBN 0 19 437003 8

First published 1990
Fourth impression 1997

Typeset in 11 on 12pt Sabon by Pentacor Ltd, High Wycombe, Bucks

This edition of *Fundamental Considerations in Language Testing*, originally published in 1990, is published by arrangement with Oxford University Press.

本书由牛津大学出版社授权上海外语教育出版社出版。

出 版 前 言

本书是讨论英语测试问题的学术专著，作者L·巴奇曼是加利福尼亚大学应用语言学和香港中文大学英语语言教学专业的教授。《TESOL季刊》和《TESOL期刊》对本书有如下评论：这本书对于从事TESOL(Teaching English to Speakers of Other Languages)的教师和学生有重要的价值；这本书是关于衡量语言能力方面的著作的极其重要的补充。

本书有三个主要目标：(1)提供有关语言测试发展和运用的实际问题的理论基础；(2)探讨语言测试的关键性问题：语言既是测试工具，又是测试的对象；(3)把语言测试和实际情景紧密联系起来。就此，本书详细论述了存在于实际语言测试发展和运用中的基本因素：衡量的特性、决定语言测试使用的情景、被衡量的语言能力的特点、衡量语言能力的测试方法的特点。

本书对意义相近的一些测试术语作了界定，分析了相互之间的关系，介绍了衡量标准的不同类型及其各自的特性，并重点介绍了衡量测试结果的两个重要特性：可靠性和有效性，以及衡量的局限性。为了将其局限性的影响降至最小，并在最大程度上提高其可靠性和有效性，作者提出了设计试题时的一系列步骤，讨论了教育项目中对语言测试的各种应用，并举出不同类型的教育项目实例，同时还扼要阐述了语言测试研究的用途，也对语言测试进行了分类。

作者为影响测试行为的两个主要方面构建出理论框架。第一个方面为被衡量者的语言能力，作者把语言交际能力分为语言能力和策略能力，并简述了语言运用中的心理、生理机制；第二个方面为测试方法的特征，这些特征涉及测试过程的一些内容，即测试环境、试题说明、应试者接收到的信息以及对此信息的预期反应。作者认为这一理

论框架不仅可用来描述现有语言测试的特征，而且可用来发展新的语言测试。不仅如此，它还可作为检验语言测试可靠性和有效性的出发点。

为了更准确地评估测试分数的可靠性以及分析测试分数时衡量错误的潜在根源，作者提出了三个衡量理论：传统的正确分数衡量理论、可归纳性理论和特定测试项反应理论（又称潜在特性理论）。作者认为测试的可靠性是其有效性的主要前提，有效性（效度）是一个单一性概念，只同特定测试运用有关。测试运用的有效性的依据可分为内容实用性（内容效度），标准关联性（标准效度）和理念效度。另外，作者还阐述了测试偏见的问题，它涉及文化、测试内容、应试者的个性、性别、年龄等许多因素，在此基础上还对测试效度在教育系统和整个社会中的道德基础问题提出了看法。

接下来，作者探讨了语言测试中有争论的问题，并为未来语言测试研究与发展的前景作了展望。在介绍语言测试本身固有的一些弊端时，指出了一些可能的解决途径，为继续研究和发展语言测试的理论提供了基础。最后，作者把语言测试作为一种职业进行了自省分析，充分地展示出他对语言测试发展的乐观态度。

本书是一本具有权威性和启发性的语言测试专著。它适用于攻读应用语言学和外语教学理论的硕士生和博士生，同时也适用于专门从事语言测试发展和使用的工作者，作为他们的教学用书和参考书。

本社编辑部

For my closest friends and loved ones: Nida, Tina, and Melissa

Acknowledgments

The author and publishers would like to thank the following for permission to reproduce the material below that falls within their copyright:

The American Council on Education for the extracts from R. L. Thorndike (ed.): *Educational Measurement* (Second edition) and R. L. Linn (ed.): *Educational Measurement* (Third edition)

The American Psychological Association for the extract from the paper by S. A. Messick in *American Psychologist* 30 .

Brooks/Cole Publishing Company for the table on page 30, adapted from M. J. Allen and W. M. Yen: *Introduction to Measurement Theory*

The Center for Applied Linguistics for the extract from the paper by J. B. Carroll in *Testing the English Proficiency of Foreign Students*

Educational and Psychological Measurement for the extracts from the paper by C. I. Mosier in Volume 7

Jay Haley for the extract from *Strategies of Psychotherapy*

The John Hopkins University Press for the figure on page 341, adapted from A. J. Nitko, 'Defining the criterion-referenced test' in R. A. Berk (ed.): *A Guide to Criterion-Referenced Test Construction*, and for the extract from the same paper

Language Learning, and the authors, for the table on page 196, from the paper by J. van Weeren and T. J. J. Theunissen in Volume 37

Newbury House Publishers for three figures (from J. A. Upshur, 'Context for language testing') in J. W. Oller and J. C. Richards: *Focus on the Learner*

Pergamon Press PLC for the extract from K. Johnson (ed.): *Communicative Syllabus Design and Methodology*

Psychometrika, and the authors, for the extract from the paper by K. K. Tatsuoka and M. M. Tatsuoka in Volume 52

Preface

This book has its origins in the close personal and collaborative relationship that Buzz Palmer and I have had for a good many years. We first hatched the idea of writing a book on language testing research somewhere between the 1981 'Illinois' study and the 1982 'Utah' study, at a time when we were both heavily committed to trying our best to incorporate what were then the still fairly new ideas about 'communicative competence' of people like Mike Canale, Merrill Swain, and Sandy Savignon into actual language tests, and to trying to find out if they were different from the kinds of language tests that were then most commonly used. The two studies that Buzz and I conducted together were a lot of hard work (neither of us may ever want to do another multitrait–multimethod study again!), but they provided a wealth of example tests and anecdotes that I have used with my classes of language testing students, and which also hopefully add a touch of both reality and comic relief to this book. More importantly, however, those studies forced us to face head-on some of the issues and problems that are the substance of this book, and to realize that addressing these will require the best ideas and tools that both applied linguistics and psychometrics have to offer. Buzz has provided me with frequent comments and suggestions as the book has taken form and he must share the credit for the inspiration and many of the ideas herein.

Much of what is in this book can also be traced to two individuals whose work has influenced my research interests, and indeed my career, in very fundamental ways. My first introduction to applied linguistics was Robert Lado's (1957) *Linguistics Across Cultures*, which was required reading for ESL Peace Corps volunteers in the mid-1960s. Even though this book was quite an eye-opener for a medieval English literature major during Peace Corps training, it wasn't until I was 'in the field', teaching ESL in a high school in the Philippines, that I began to appreciate its wisdom. Its real impact on my career, however, came a few years later, when I was drawn back to it, during graduate school, after having read John B. Carroll's

(1964) *Language and Thought.* It was Carroll's discussions of language acquisition research and cross-cultural research in psycholinguistics, along with Lado's discussion of contrasts across languages, that I found both exciting and challenging, and that piqued an interest that eventually led me to abandon medieval literary studies for dissertation research in second language acquisition.

It was not until after graduate school, when, as a Ford Foundation 'adviser', I found myself in charge of the development and administration of language tests at a national language center in Thailand, that my on-the-job learning led me to the library, where I first discovered that either Lado or Carroll had anything to do with language testing! During the next few years I was fortunate to have the opportunity to work with John Carroll on several occasions, on the development of language aptitude tests in Thai, and was always both awed and inspired by his encyclopedic knowledge, his brilliant insights, and his consummate craftsmanship. I continue to read his work with interest and to correspond with him on occasion to ask a question or pose a problem for his consideration. A great deal of whatever is useful in this book is a result of my contact with him and his work.

When I was trying to come up with a title for this book, it seemed that all the good titles had already been taken. There have been titles in language testing with 'issues' (for example, Oller 1983b; Alderson and Hughes 1981; Lowe and Stansfield 1988), 'current developments' (Hughes and Porter 1983), 'problems' (Upshur and Fata 1968; Interuniversitäre Sprachtestgruppe Symposium Proceedings: Culhane *et al.* 1981, 1984; Klein-Braley and Stevenson 1981; Kohonen *et al.* 1985; Lutjeharms and Culhane 1982), 'approaches' (Spolsky 1978a; Brindley 1986), 'directions' (Read 1981; Lee *et al.* 1985), 'concepts' (Brière and Hinofotis 1979a) and 'research' (Oller and Perkins 1980; Oller 1983b; Bailey *et al.* 1987). And while I'm not aware of any 'principles' or 'essentials' titles in language testing, I'm not convinced that what I have to offer is quite as certain as these terms would imply. The title I've chosen turns out to be a portmanteau of the titles of two seminal works in language testing that happen to have been published in the same year: 'Fundamental considerations in the testing for English language proficiency of foreign students' (Carroll 1961a) and *Language Testing* (Lado 1961). Thus, in solving my title problem, I also echo my debt to Lado and Carroll; hopefully what I've taken from them is returned in some small measure in the pages that follow.

Throughout the travail of writing this book, I have (sometimes)

heeded the counsel, or head-bashing, if you will, of a group of individuals who have been my severest critics, and who have also aided and abetted me in this endeavor. Their written comments on various versions and parts of the manuscript have both kept me clearly attuned to fundamental issues, and pushed me to discuss areas that I might have wanted to avoid. They must therefore rightfully share the credit for what is good, and take their lumps as co-conspirators for whatever errors there are that came from them. Among those that should be thus implicated are Charles Alderson, Doug Brown, J. D. Brown, Larry Bouton, Gary Buck, Mike Canale, Gary Cziko, Fred Davidson, John de Jong, Antony Kunnan, Brian Lynch, John Oller, Sandy Savignon, Larry Selinker, Bernard Spolsky, Jack Upshur, and Swathi Vanniarajan. Comments from Gillian Brown on Chapters 4 and 5 were also very helpful. I am most grateful to Charles Alderson, John Carroll, John Clark, Bernard Spolsky, and Henry Widdowson, whose meticulous reading of the manuscript and insightful comments, from different perspectives, have improved it immensely. I would particularly like to thank Yukiko Abe-Hatasa, Buzz Palmer, Larry Selinker, and Jack Upshur for their comments and suggestions, based on their use of the book in manuscript form with their classes on language testing, and Sasi Jungsatitkul, who helped write the discussion questions. Finally, my sincerest gratitude goes to my own students, whose insights, questions, and comments have led me to sharpen my thinking on many issues, and to recognize (and admit) where I remain fuzzy and uncertain. I thank them also for patiently bearing the burden of helping me refine my presentation of these issues.

Writing this book has been challenging and rewarding in that it has given me the opportunity to work my way through some of the conundrums of language testing and to reach, if not solutions, at least a sense of direction and a strategy for research. It has also been a source of frustration, however, as I see the field moving at a pace beyond my ability to incorporate developments into the present discussion. Even as I write this preface, for example, I have received the manuscript of a 'state of the art' article on language testing from Peter Skehan, and from Liz Hamp-Lyons a review article of recent and forthcoming textbooks in applied linguistics research and language testing. These articles review recent work in language testing, and relate this to research in other areas of applied linguistics. Also in my mail is the list of titles of papers for the upcoming 11th Annual Language Testing Research Colloquium, which promise to report recent developments in a number of areas.

But while these developments may be a source of minor frustration to me, as I attempt to reach closure on this book, at the same time they give me cause for optimism. Language testers now have their own journal, *Language Testing*; three newsletters, *Language Testing Update*, the *AILA Language Testing News*, and the *IATEFL Testing SIG Newsletter*, and can count at least three major international conferences annually (the Language Testing Research Colloquium (LTRC) in North America, the Interuniversitäre Sprachtestgruppe (IUS) Symposium in Europe, and the Academic Committee for Research on Language Testing (ACROLT) Symposium in Israel), as well as several regional conferences, such as those in Estonia, Japan, and Thailand, which regularly focus on issues in language testing. What is most encouraging about these events and developments is that the concerns of language testing are drawing together a widening circle of applied linguists, language teachers, and psychometricians, who recognize the interrelatedness of their needs, interests, and areas of expertise, and whose collaboration can only advance our understanding of language ability and how we can most effectively and usefully measure it.

Savoy, Illinois
February 1989

Contents

1 Introduction

The aims of the book

In developing and using measures of language abilities, we are constantly faced with practical questions, 'What type of test should we use?', 'How long should the test be?', 'How many tests do we need to develop?', questions to which there are no clear-cut, absolute answers. Other questions are even more difficult to answer. For example, 'How reliable should our test be?', 'Are our test scores valid for this use?', and 'How can we best interpret the results of our test?' In addressing questions such as these, we inevitably discover that the answers depend upon a wide range of prior considerations. Since these considerations will vary from one test context to the next, an appropriate answer for one situation may be inappropriate for another. Thus, in developing and using language tests we are seldom, if ever, faced with questions to which there are right or wrong answers. Answering these questions always requires consideration of the specific uses for which the test is intended, how the results are to be interpreted and used, and the conditions under which it will be given.

This book is not a 'nuts and bolts' text on how to write language tests. Rather, it is a discussion of fundamental issues that must be addressed at the start of any language testing effort, whether this involves the development of new tests or the selection of existing tests. How we conceive of these issues will affect how we interpret and use the results of language tests. One objective of this book is thus to provide a conceptual foundation for answering practical questions regarding the development and use of language tests. This foundation includes three broad areas: (1) the context that determines the uses of language tests; (2) the nature of the language abilities we want to measure, and (3) the nature of measurement. This conceptual foundation is applicable to a wide range of general concerns in language testing, including diagnostic, achievement, and language aptitude testing. Furthermore, this foundation provides a

basis for addressing issues in the measurement of language proficiency, which presents some of the most complex and challenging problems for language testing, problems to which much of the discussion of this text is addressed.

A second objective of this book is to explore some of the problems raised by what is perhaps a unique characteristic of language tests and a dilemma for language testers – that language is both the instrument and the object of measurement – and to begin to develop a conceptual framework that I believe will eventually lead, if not to their solution, at least to a better understanding of the factors that affect performance on language tests. Unlike tests of other abilities or areas of knowledge, where we frequently use language in the process of measuring something else, in language tests, we use language to measure language ability. What I believe this means is that many characteristics of the instrument, or the method of observing and measuring, will overlap with characteristics of the language abilities we want to measure. In order to understand how these characteristics interact, as I believe they do, and how they affect performance on language tests, I believe we must develop a framework for describing the characteristics of both the language abilities we want to measure and of the methods we use to measure these abilities.

The climate for language testing

Language testing almost never takes place in isolation. It is done for a particular purpose and in a specific context. A third objective of this book is thus to relate language testing to the contexts in which it takes place. Current research and development in language testing incorporates advances in several areas: research in language acquisition and language teaching, theoretical frameworks for describing language proficiency and language use, and measurement theory.[1]

Research in language acquisition and language teaching

As Upshur (1971) noted several years ago, there is an intrinsic reciprocal relationship between research in language acquisition and developments in language teaching on the one hand, and language testing on the other. That is, language testing both serves and is served by research in language acquisition and language teaching. Language tests, for example, are frequently used as criterion measures of language abilities in second language acquisition research. Similarly, language tests can be valuable sources of