Statistical Analyses for Language Assessment

Lyle F. Bachman

Cambridge Language Assessment Series

Series Editors J. Charles Alderson & Lyle F. Bachman

CAMBRIDGE UNIVERSITY PRESS

Statistical Analyses for Language Assessment

Lyle F. Bachman

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2004

Printed in the United Kingdom at the University Press, Cambridge

Typeface 9.5/13pt Utopia. *System* QuarkXPress™ [SE]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

ISBN 0 521 80277 6 hardback
ISBN 0 521 00328 8 paperback

Statistical Analyses for Language Assessment

THE CAMBRIDGE LANGUAGE ASSESSMENT SERIES

Series editors: J. Charles Alderson and Lyle F. Bachman

In this series:
Assessing Grammar by James E. Purpura
Assessing Languages for Specific Purposes by Dan Douglas
Assessing Listening by Gary Buck
Assessing Reading by J. Charles Alderson
Assessing Speaking by Sari Luoma
Assessing Vocabulary by John Read
Assessing Writing by Sara Cushing Weigle

To Nida.

Series Editor's Preface

Language testing is frequently associated with statistics in the mind of language educators. Often for that very reason, language teachers feel that language testing is a somewhat alien discipline. Language teachers have indeed often chosen their career in reaction against the scientific disciplines they experienced in their school years, feeling that the Arts and Humanities are more in tune with their own interests, inclinations and competences.

The Cambridge Language Assessment Series (CLAS) does not share that view of language testing, as all the volumes published in the series to date attest. The CLAS view of language testing is that what is central to the discipline is language: the constructs that the tests and assessment procedures are seeking to evaluate. A solid background in applied linguistics or a related discipline is in our view crucial for the development of appropriate measures. However, the reader should note the use of the word 'measure'. Language tests are intended to measure, and without quantification they cannot measure. Quantification implies numbers and numbers imply statistics. And so, although a firm understanding of the nature of language is essential for the trained language tester, so too is at least a basic familiarity with statistics, that is, those statistics that are commonly used in analysis and investigation of and with language tests.

This volume thus aims to complement the other volumes in CLAS by providing a firm grounding in such statistical procedures as are needed for language test analysis. There are, of course, many introductory textbooks to statistics, but there are very few that are specifically focussed on the needs of language testers. Yet it is common knowledge that one learns best

by applying the theory one is learning to the solution of practical problems within one's area of experience, and so the best way for language testers to learn statistics is in the context of language tests. This book does precisely that: the examples that are used to illustrate and explain are taken from the field of language testing. And Lyle Bachman is himself an internationally known and respected applied linguist and language testing researcher. His almost unique combination of applied linguistic and statistical expertise makes him very well qualified to author this volume, and the authority with which he writes carries weight and conviction.

An overriding consideration in designing, developing and using language tests is that of usefulness. Bachman and Palmer (1996) define test usefulness as consisting of six qualities: reliability, construct validity, authenticity, interactiveness, impact and practicality. An important part of their approach is what they call the logical, or subjective, evaluation of usefulness by the test developer during the design and development of the test.

However, the usefulness of a test also depends to a great extent on how test takers perform on that test. The evaluation of test usefulness must therefore also include the empirical investigation of test performance. One crucial aspect of this performance is the scores that the test takers obtain. Test developers must go beyond mere assertions of usefulness, and provide *evidence* that supports the claims they make about how test scores can be appropriately interpreted and used. Similarly, it is the responsibility of test users to require test developers to provide such evidence, and to use this evidence appropriately and ethically in their selection and use of language tests.

The primary purpose of this book is to make the knowledge and skills – the 'tools' – of statistical analysis accessible to classroom teachers and other applied linguists who may need to develop and use language tests. Statistical analyses provide ways to corroborate, with empirical evidence, beliefs and claims about the tests we develop or use. Statistical analyses can also show where tests fall short of expectations, and provide guidance for improving them. For these reasons, the tools of statistical analysis are essential to any test development and use.

The presentations in this book focus on the rationales and appropriate applications for the use of these procedures for language testing, rather than on theoretical or mathematical exposition. For each of the procedures discussed, Bachman provides the following:

1 a non-technical rationale for using the procedure, including the type of information the procedure will provide and what we can do with this information;

2 a discussion of the conditions under which the procedure may be used appropriately, and where there is a choice of more than one procedure, considerations for making an informed choice;

3 explicit step-by-step instructions for the procedure; and

4 guidelines for using the results of the procedure.

The book is in three parts. Part I discusses terms, concepts and statistical procedures that are basic to the analyses presented in the later chapters, Part II discusses statistical procedures for analyzing test scores for the purposes of improving their usefulness, and Part III discusses statistical procedures that can inform the ways in which we interpret and use test scores. Readers who are unfamiliar with quantitative methodology may want to read the chapters in the order presented. Readers who are already familiar with the basic concepts may want to go directly to those chapters that are most relevant to their particular needs and interests.

However, statistical concepts and associated procedures are best learned by doing. It is impossible to learn all the necessary procedures and apply them accurately and appropriately without considerable and repeated practice. Perhaps the strongest feature of this book, in addition to the fact that the text itself is excellent – clear, coherent, practical, user friendly and of an appropriate length – is the fact that it is accompanied by a practice Workbook and a CD containing data for statistical analysis.

No textbook on statistics can possibly familiarize readers and users with everything they need to know in order to analyze language tests adequately. What is needed is constant practice and feedback.

Learning how to use the statistical procedures discussed in this book will thus be greatly facilitated by practice in applying them and working them through with the actual data from language tests included on the CD. The Workbook includes extensive practice exercises with these data sets. The exercises provide opportunities for readers to apply what they have learned from this book in calculating the appropriate statistical procedures and interpreting the results. The exercises provide practice in hand calculations with small data sets and in the use of the SPSS computer program (SPSS Inc., 2002) for larger data sets.

Anyone using this book and the associated Workbook and CD should be able to become comfortable with and confident in his or her ability to use statistical analyses appropriately in their own work, whether this be language testing or other areas of applied linguistics research.

J Charles Alderson

Acknowledgements

I would first like to thank the many students who have struggled with me through several drafts of this book. Their comments, questions, groans, and flashes of understanding and insight have shaped virtually every sentence in this book. These are the folks who have kept my writing alive these past few years and who have provided me with a touchstone of reality; their classes were the crucible in which this book was shaped.

I would also like to thank my friend, colleague and fellow editor of many years, Charles Alderson, for giving me the kind of sustained support I needed while writing this book. He knew when to nudge me and when to leave me alone. His comments and suggestions at times helped keep me on track, and at other times pushed me in directions I hadn't thought of or hadn't wanted to go. He was always constructive, but never let me take the easy path, and for this I thank him.

I would like to thank our editor at CUP, Mickey Bonin, who was a wonderful support throughout the process of transforming an idea and disciplining it into published form.

Finally, I would like to thank several anonymous reviewers whose comments and suggestions on the proposal helped shape the book in its initial stages; Lynn Winters for her helpful comments and suggestions on earlier drafts of the first two chapters, and to Antony Kunnan, Craig Deville, Yasuyo Sawaki and Xiaoming Xi for their insightful comments and suggestions on the penultimate draft of the book. I would also like to thank Jeong-won Lee for her careful proofreading of the page proofs. These reviewers are responsible for urging me to include much that I had not originally planned to include, and the content of the book is, I believe, much better for this. I alone can be held responsible for the felicitousness or lack thereof in the presentation of the content.

Abbreviations

CAT	computer adaptive testing
CFA	confirmatory factor analysis
CI	confidence interval
CLA	communicative language ability
CR	criterion-referenced
CTT	classical test theory
EFA	exploratory factor analysis
G-theory	generalizability theory
IA	item analysis
ICC	item characteristic curve
IIF	item information function
IRT	item response theory
L1	first language
L2	second language
MFRM	many-facet Rasch measurement
MTMM	multitrait–multimethod correlation matrix
NR	norm-referenced
SEM	standard error measurement
TIF	test information function
TLU	target language use
TMF	test method facets

Contents

PART I

Basic concepts and statistics

Statistical analyses provide a set of tools for helping us to evaluate and improve the qualities of the tests we use, and to help us assure that we use these tests in ways that are valid and fair. The appropriate use of statistical procedures rests on an understanding of basic concepts and logic, as well as calculations of the statistics themselves. The first part of this book discusses the conceptual bases and logical foundations of language testing and measurement, as well as the basic statistical procedures that provide the foundation for the rest of the book.

Chapter 1 discusses the applied linguistic and measurement contexts that constitute the joint foundations of language assessment. Chapter 2 discusses procedures for describing the statistical properties of a set of test scores. Chapter 3 discusses procedures for calculating correlation coefficients, which can be used to investigate the relationship between two sets of test scores.

CHAPTER ONE

Basic concepts and terms

Language tests have become a pervasive part of our education system and society. Scores from language tests are used to make inferences about individuals' language ability and to inform decisions we make about those individuals. For example, we use language tests to help us identify second or foreign language learners in schools, to select students for admission to universities, to place students into language programs, to screen potential immigrants and to select employees. Language tests thus have the potential for helping us collect useful information that will benefit a wide variety of individuals. However, to realize this potential, we need to be able to demonstrate that scores we obtain from language tests are reliable, and that the ways in which we interpret and use language test scores are valid. If the language tests we use do not provide reliable information, and if the uses we make of these test scores cannot be supported with credible evidence, then we risk making incorrect and unfair decisions that will be potentially harmful to the very individuals we hope to benefit. Thus, if we want to assure that we use language tests appropriately, we need to provide evidence that supports this use. An important kind of evidence that we collect to support test use is that which we derive from quantitative data – scores from test tasks and tests as a whole – and the appropriate statistical analyses of these data. An understanding of the nature of quantitative data and how to analyze these statistically is thus an essential part of language testing.

Much of the data we obtain from language assessment is quantitative, consisting of numbers, and **statistics** is a set of logical and mathematical procedures for analyzing quantitative data. In order to appropriately use