

# Practical Language Testing

Glenn Fulcher

The background of the cover is a monochromatic blue with a high-speed photograph of water droplets hitting a surface, creating ripples and splashes. The droplets are captured in various stages of impact, with some showing a crown-like splash and others as a vertical column of water.

ROUTLEDGE

The Routledge logo, which consists of a stylized white 'R' shape.

# Practical Language Testing

Glenn Fulcher

For all the inspiring teachers  
I have been lucky enough to have  
and especially  
Revd Ian Robins  
Who knows where the ripples end?

First published in Great Britain in 2010 by  
Hodder Education, An Hachette UK Company,  
338 Euston Road, London NW1 3BH

© 2010 Glenn Fulcher

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronically or mechanically, including photocopying, recording or any information storage or retrieval system, without either prior permission in writing from the publisher or a licence permitting restricted copying. In the United Kingdom such licences are issued by the Copyright Licensing Agency: Saffron House, 6–10 Kirby Street, London EC1N 8TS.

Hachette UK's policy is to use papers that are natural, renewable and recyclable products and made from wood grown in sustainable forests. The logging and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

The advice and information in this book are believed to be true and accurate at the date of going to press, but neither the author nor the publisher can accept any legal responsibility or liability for any errors or omissions.

*British Library Cataloguing in Publication Data*

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

*Library of Congress Cataloging-in-Publication Data*

A catalog record for this book is available from the Library of Congress

ISBN: 978 0 340 984482

1 2 3 4 5 6 7 8 9 10

Cover Image © Anthony Bradshaw/Photographer's Choice RF/Getty Images  
Typeset in 10 on 13pt Minion by Phoenix Photosetting, Chatham, Kent  
Printed and bound in Great Britain by Antony Rowe, Chippenham, Wilts

What do you think about this book? Or any other Hodder Education title? Please send your comments to [educationenquiries@hodder.co.uk](mailto:educationenquiries@hodder.co.uk)

<http://www.hoddereducation.com>



# Contents

<i>Acknowledgements</i>	vii
<i>List of figures</i>	ix
<i>List of tables</i>	xi
<i>Preface</i>	xiii
<b>1 Testing and assessment in context</b>	<b>1</b>
1. Test purpose	1
2. Tests in educational systems	4
3. Testing rituals	5
4. Unintended consequences	6
5. Testing and society	8
6. Historical interlude I	11
7. The politics of language testing	12
8. Historical interlude II	15
9. Professionalising language education and testing	17
10. Validity	19
Activities	21
<b>2 Standardised testing</b>	<b>31</b>
1. Two paradigms	31
2. Testing as science	32
3. What's in a curve?	35
4. The curve and score meaning	36
5. Putting it into practice	37
6. Test scores in a consumer age	42
7. Testing the test	44
8. Introducing reliability	46
9. Calculating reliability	47
10. Living with uncertainty	54
11. Reliability and test length	57
12. Relationships with other measures	57
13. Measurement	59
Activities	60

<b>3 Classroom assessment</b>	67
1. Life at the chalk-face	67
2. Assessment for Learning	68
3. Self- and peer-assessment	70
4. Dynamic Assessment	72
5. Understanding change	75
6. Assessment and second language acquisition	77
7. Criterion-referenced testing	79
8. Dependability	81
9. Some thoughts on theory	87
Activities	90
<b>4 Deciding what to test</b>	93
1. The test design cycle	93
2. Construct definition	96
3. Where do constructs come from?	102
4. Models of communicative competence	105
5. From definition to design	118
Activities	120
<b>5 Designing test specifications</b>	127
1. What are test specifications?	127
2. Specifications for testing and teaching	134
3. A sample detailed specification for a reading test	139
4. Granularity	147
5. Performance conditions	148
6. Target language use domain analysis	149
7. Moving back and forth	154
Activities	155
<b>6 Evaluating, prototyping and piloting</b>	159
1. Investigating usefulness and usability	159
2. Evaluating items, tasks and specifications	159
3. Guidelines for multiple-choice items	172
4. Prototyping	173
5. Piloting	179
6. Field testing	185
7. Item shells	186
8. Operational item review and pre-testing	188
Activities	190
<b>7 Scoring language tests</b>	197
1. Scoring items	197

2. Scorableity	201
3. Scoring constructed response tasks	208
4. Automated scoring	216
5. Corrections for guessing	218
6. Avoiding own goals	219
Activities	220

## **8 Aligning tests to standards** 225


1. It's as old as the hills	225
2. The definition of 'standards'	225
3. The uses of standards	226
4. Unintended consequences revisited	228
5. Using standards for harmonisation and identity	229
6. How many standards can we afford?	231
7. Performance level descriptors (PLDs) and test scores	233
8. Some initial decisions	234
9. Standard-setting methodologies	236
10. Evaluating standard setting	241
11. Training	243
12. The special case of the CEFR	244
13. You can always count on uncertainty	248
Activities	250

## **9 Test administration** 253

1. No, no. Not me!	253
2. Controlling extraneous variables	254
3. Rituals revisited	258
4. Standardised conditions and training	259
5. Planned variation: accommodations	262
6. Unplanned variation: cheating	264
7. Scoring and moderation	267
8. Data handling and policy	268
9. Reporting outcomes to stakeholders	269
10. The expense of it all	272
Activities	274

## **10 Testing and teaching** 277

1. The things we do for tests	277
2. Washback	277
3. Washback and content alignment	282
4. Preparing learners for tests	288
5. Selecting and using tests	292
6. The gold standard	295



vi Contents

Activities	298
<i>Epilogue</i>	300
<i>Appendices</i>	301
<i>Glossary</i>	319
<i>References</i>	325
<i>Index</i>	343

# Acknowledgements

I am deeply indebted to the Leverhulme Trust ([www.leverhulme.ac.uk](http://www.leverhulme.ac.uk)), which awarded me a Research Fellowship in 2009 in order to carry out the research required for this book, and funded study leave to write it. The generosity of the Trust provided the time and space for clear thinking that work on a text like this requires.

The University of Leicester was extremely supportive of this project, granting me six months' study leave to work entirely on the book. I would also like to thank staff in the School of Education for help and advice received while drafting proposals and work schedules.

I am grateful to the people, and the institutions, who have given me permission to use materials for the book.

Special thanks are due to Professor Yin Jan of Shanghai Jiao Tong University, and Chair of the National College English Testing Committee of the China Higher Education Department. Her kindness in providing information about language testing in China, as well as samples of released tests, has enriched this book.

I have always been inspired by my students. While I was working on the development of Performance Decision Trees (see Chapter 7), Samantha Mills was working on a dissertation in which she developed and prototyped a task for use in assessing service encounter communication in the tourist industry. In this book the two come together to illustrate how specifications, tasks and scoring systems, can be designed for specific purpose assessment. I am very grateful to Samantha for permission to reproduce sections of her work, particularly in Chapters 5 and 6.

Test design workshops can be great fun; and they are essential when brainstorming new item types. I have run many workshops of this kind, and the material used to illustrate the process of item evaluation in Chapter 6 is taken from a workshop conducted for Oxford University Press (OUP). I am grateful to OUP, particularly Simon Beeston and Alexandra Miller, for permission to use what is normally considered to be confidential data.

The book presents a number of statistical tools that the reader can use when designing or evaluating tests. All of the statistics can be calculated using packages such as SPSS, or online web-based calculators. However, I believe that it is important for people who are involved in language testing to understand how the basic statistics can be calculated by hand. My own initial statistical training was provided by Charles Owen at the University of Birmingham, and I have always been grateful that he made us do calculations by hand so that we could 'see' what the machine was doing. However, calculation by hand can always lead to errors. After a while, the examples in the text became so familiar that I would not have been able to spot any errors, no matter how glaring. I am therefore extremely grateful to Sun Joo Chung of the University of Illinois at Urbana-Champaign for the care with which she checked and corrected these parts of the book.



The content of the book evolved over the period during which it was written. This is because it is based on a research project to discover the language testing needs of teachers and students of language testing on applied linguistics programmes. A survey instrument was designed and piloted, and then used in the main study. It was delivered through the Language Testing Resources website (<http://languagetesting.info>), and announced on the language testing and applied linguistics discussion lists. It was also supported by the United Kingdom's Subject Centre for Languages, Linguistics and Area Studies. The respondents came from all over the world, and from many different backgrounds. Each had a particular need, but common themes emerged in what they wished to see in a book on practical language testing. The information and advice that they provided has shaped the text in many ways, as my writing responded to incoming data. My thanks, therefore, to all the people who visited my website and spent time completing the survey.

My thanks are also due to Fred Davidson, for a continued conversation on language testing that never fails to inspire. To Alan Davies and Bernard Spolsky, for their help and support; and for the constant reminder that historical context is more important than ever to understanding the 'big picture'. And to all my other friends and colleagues in the International Language Testing Association (ILTA), who are dedicated to improving language testing practice, and language testing literacy.

Every effort has been made to obtain the necessary permission with reference to copyright material. The publishers apologise if inadvertently any sources remain unacknowledged and will be glad to make the necessary arrangements at the earliest opportunity.

Finally, acknowledgements are never complete with recognition for people who have to suffer the inevitable lack of attention that writing a book generates. Not to mention the narrowing of conversational topics. My enduring thanks to Jenny and Greg for their tolerance and encouragement.

# Figures

- 1.1 Jeremy Bentham's Panopticon in action
- 2.1 Distribution of scores in typical army groups, showing value of tests in identification of officer material
- 2.2 The curve of normal distribution and the percentage of scores expected between each standard deviation
- 2.3 A histogram of scores
- 2.4 The curve of normal distribution with raw scores for a particular test
- 2.5 The curve of normal distribution with the meaning of a particular raw score
- 2.6 A scatterplot of scores on two administrations of a test
- 2.7 Shared variance between two tests at  $r^2 = .76$
- 2.8 Confidence intervals
- 3.1 Continuous assessment card
- 3.2 An item from an aptitude test
- 3.3 A negatively skewed distribution
- 4.1 The test design cycle
- 4.2 The levels of architectural documentation
- 4.3 Language, culture and the individual
- 4.4 Canale's expanded model of communicative competence
- 4.5 Bachman's components of language competence
- 4.6 The common reference levels: global scale
- 5.1 Forms and versions
- 5.2 Popham's (1978) five-component test specification format
- 7.1 Marking scripts in 1917
- 7.2 The IBM 805 multiple-choice scoring machine
- 7.3 Example of a branching routine
- 7.4 An Item-person distribution map
- 7.5 EBB for communicative effectiveness in a story retell
- 7.6 A performance decision tree for a travel agency service encounter
- 8.1 The distributions of three groups of test takers
- 9.1 An interlocutor frame
- 10.1 An observation schedule for writing classes

*This page intentionally left blank*

# Tables

- 2.1 Deviation scores
- 2.2 Proportion of test takers from two groups answering individual items correctly
- 2.3 Calculating a correlation coefficient between two sets of scores
- 2.4 Item variances for the linguality test
- 2.5 Descriptive statistics for two raters, rating ten essays
- 2.6 Descriptive statistics for combined scores
- 2.7 Correlations of group with individual linguality test scores
- 2.8 The relation between the two tests
- 3.1 A classification table
- 3.2 Results of a reading test
- 6.1 Distractor analysis
- 6.2 Responses of 30 students to items 67–74
- 6.3 Standard deviation
- 6.4 Means for p and q for item 70
- 7.1 Correlations between human and machine scores on PhonePass SET-10
- 8.1 A truth table
- 8.2 Classifications of students into three levels by two judges
- 9.1 Observed values by conditions and outcomes on a language test
- 9.2 Expected values by outcomes on a language test
- 9.3 Critical values of chi-square
- 10.1 Standards for formative writing, language arts, grades 9–12
- 10.2 Standards for summative writing, language arts, grades 9–12

*This page intentionally left blank*

# Preface

This book is about building and using language tests and assessments. It does what it says on the tin: it is a *practical* approach. However, it does not provide ready-made solutions. Language testing is a complex social phenomenon, and its practice changes lives. The book therefore assumes that you will wish to think carefully about testing and its impact in your own context.

The term ‘practical’ therefore needs some definition. The book is ‘practical’ in the sense that it gives guidance on how to do things to build a test. It is also ‘practical’ in that each chapter will be useful to you when you come to making decisions about when, why and how to conduct assessments. The book is designed to provide the *knowledge* you will need to apply, and the *skills* you will need to practise. However, if we are to build good language tests, we have to be aware of the larger social, ethical, and historical context, within which we work. If language testing and assessment are not guided by *principles*, we could end up doing more harm than good. Davies (2008a) has cogently argued that testing and assessment texts that do not embed knowledge and skills in principles ignore the increasing demand of professionalism and social responsibility.

Language professionals, applied linguists and educational policy makers need an expanded ‘assessment literacy’ in order to make the right decisions for language learners and institutions (Taylor, 2009). This literacy will be about learning the nuts and bolts of writing better test items (Coniam, 2008), and establishing a core knowledge base (Inbar, 2008); but it is also about appreciating the reasons why we test, why we test the way we do and how test use can enrich or destroy people’s hopes, ambitions and lives.

Although I am far from being in the ‘postmodern’ school of language testing and assessment, the view that language testing is a social activity cannot be denied (McNamara, 2001). Nor can the fact that our practices are thoroughly grounded in a long history that has brought us to where we are (Spolsky, 1995). It is partly because of this history that many texts published ‘for teachers’ focus almost entirely upon the technologies of normative large-scale standardised testing. While it is important that teachers are familiar with these, they are not always directly relevant to the classroom. This book therefore tries to introduce a balance between standardised testing and classroom assessment.

The structure reflects a conscious decision to place language testing and assessment within context, *and* to provide the ‘practical’ guidance on the nuts and bolts of test building. Broadly, the first three chapters survey the language testing landscape upon which we can build. Chapter 4 is about the material that we can use in construction, and the rest of the book takes the reader through the process of building and implementing a language test.

Chapter 1 considers the purpose of testing in the broadest sense of why societies use tests, and in the narrow sense of how we define the purpose of a particular test. It looks

at how tests are used, for good and ill; and the unintended consequences that testing can have on people who are caught up in the need to succeed. Chapters 2 and 3 deal in turn with large-scale standardised testing, and then with classroom assessment. The stories of both paradigms are set within a historical framework so that you can see where the theories and practices originate.

In Chapter 4 we begin the journey through the process of test design, starting with deciding what to test, and why. Chapter 5 begins the test design process in earnest, where we discuss how to create test specifications – the basic design documents that help us to build a test. This is where we learn to become ‘test architects’, shaping the materials and putting them together in plans that can be used to produce usable test forms. In Chapter 6 we look at how to evaluate the test specifications and test items, from initial critical discussions in specification workshops to trying out items and tests with learners. Chapter 7 contains a discussion of scoring, covering both traditional item types like multiple choice, as well as performance tests that require human judgement. Frequently, we have to use tests to make decisions that require a ‘cut score’ – a level on the test above which a test taker is judged to be a ‘master’, and below which they are still ‘novices’. Establishing cut scores and linking these to absolute standards is the subject of Chapter 8. Chapter 9 discusses the practicalities of test administration, and why the ‘rituals’ of testing have grown as they have.

Finally, in Chapter 10, we return to the classroom and to the effect that tests have upon learning and teaching, and how we go about preparing learners to take tests.

Throughout the book I have included examples from real tests and assessments. Some of these are good examples that we can emulate. Others are provided for you to critique and improve. Some of them are also drawn from historical sources, as ‘distance’ is useful for nurturing critical awareness. However, I do not present sets of typical test items and tasks that you could simply select to include in your own tests. There are plenty of books on the market that do this. This book asks you to think about what item or task types would be most useful for your own tests. We discuss options, but only you can provide the answers and the rationales for the choices you make.

There are activities at the end of every chapter that you can attempt on your own, although many would benefit from team work. Sharing experiences and debating difficult issues is best done in a group. And it’s also more fun. The activities have been designed to help you think through issues raised in the chapter, and practise the skills that you have learned. The activities are not exhaustive, and you are encouraged to add to these if you are using the text in a language testing course. Beginning in Chapter 4 there is also a Project that you may wish to do as you move from chapter to chapter.

This structure has been shaped not only by my own understanding of what an introductory book to foster ‘assessment literacy’ might look like, but also by what language teachers and students of applied linguistics have told me that they need to know, and be able to do. Prior to writing the book I undertook a large-scale internet-based survey, funded by the Leverhulme Trust. Almost 300 respondents completed the survey, and I was struck by the sophistication of their awareness of assessment issues.

Here is a selection of typical responses to a question about what teachers and students of applied linguistics most need in a ‘practical’ language text:

*Evaluating reliability for our in-house tests, and checking questions at each stage in test development.*

*I don’t understand statistics, but I know they can be useful. I need it explaining conceptually.*

*We need to know the jargon, but introduce it step by step.*

*Hands-on activities; examples of test specs; a glossary would be useful.*

*A book of this type must focus on the basics of item writing and test construction, the basic concepts of validity and reliability, particularly in regards to the assessment of speaking and writing. It must also cover the ethics of test use and test score interpretation.*

*Developing classroom tests, performance tests, setting score standards, deciding what to test, preparing learners for test situations.*

*Differentiation between classroom assessments, formative assessment, and large-scale assessment when discussing key issues.*

*Most of the assessment/testing practices are done by teachers; I think that a book should be aimed at ‘normal’ language teachers more than specialists in testing, they already have other sources of information and training.*

*Issues to do with ensuring validity and reliability in language testing. The test writing process from the creation of test specifications through to the trialling, administration and marking of tests.*

*Vignettes; glossary; application activities for individuals and groups, including some practice with basic test statistics and approaches to calculating grades.*

*Some information on testing as an industry, a multi-billion dollar concern and why we have to fight crap when we see it.*

Luckily, many respondents said they realised that it is impossible to include everything in a practical language testing book. This is evidently true, as you will see. I am sure to have left out a topic that you think should have been included. One respondent understood this all too well: ‘The book should be well-structured, clearly focused, and however tempted you might be to put everything into one book, you should be selective in order to be comprehensible and user-friendly.’ I am not entirely sure that I have achieved this. But if I have got even halfway there, my time will have been well spent.

As another respondent said, ‘The learning never ends.’ In order to sustain you during your journey through the book, you may wish to pay regular visits to my website:

<http://languagetesting.info>



Here you will find a set of online videos that define and explain some of the key concepts and topics in language testing. To help you with additional reading, I have links to online articles, and other language testing websites. There are links to useful journals, and regular updates on testing stories that get into the news.

Constructive criticism is always welcome, via the website.