COMPANION WEBSITE

Language Testing and Assessment

an advanced resource book

Glenn Fulcher and Fred Davidson

ROUTLEDGE APPLIED LINGUISTICS

SERIES EDITORS: CHRISTOPHER N. CANDLIN AND RONALD CARTER

LANGUAGE TESTING AND ASSESSMENT

Routledge Applied Linguistics is a series of comprehensive resource books, providing students and researchers with the support they need for advanced study in the core areas of English language and Applied Linguistics.

Each book in the series guides readers through three main sections, enabling them to explore and develop major themes within the discipline.

- Section A, Introduction, establishes the key terms and concepts and extends readers' techniques of analysis through practical application.
- Section B, Extension, brings together influential articles, sets them in context and discusses their contribution to the field.
- Section C, Exploration, builds on knowledge gained in the first two sections, setting thoughtful tasks around further illustrative material. This enables readers to engage more actively with the subject matter and encourages them to develop their own research responses.

Throughout the book, topics are revisited, extended, interwoven and deconstructed, with the reader's understanding strengthened by tasks and follow-up questions.

Language Testing and Assessment:

- provides an innovative and thorough review of a wide variety of issues from practical details of test development to matters of controversy and ethical practice
- investigates the importance of the philosophy of pragmatism in assessment, and coins the term 'effect-driven testing'
- explores test development, data analysis, validity and their relation to test effects
- illustrates its thematic breadth in a series of exercises and tasks, such as analysis of test results, study of test revision and change, design of arguments for test validation and exploration of influences on test creation
- presents influential and seminal readings in testing and assessment by names such as Michael Canale and Merrill Swain, Michael Kane, Alan Davies, Lee Cronbach and Paul Meehl and Pamela Moss.

Written by experienced teachers and researchers in the field, *Language Testing and Assessment* is an essential resource for students and researchers of Applied Linguistics.

Glenn Fulcher is Senior Lecturer in the School of Education at the University of Leicester, UK.

Fred Davidson is Associate Professor in the Division of English as an International Language at the University of Illinois at Urbana-Champaign, USA.

ROUTLEDGE APPLIED LINGUISTICS

SERIES EDITORS

Christopher N. Candlin is Senior Research Professor in the Department of Linguistics at Macquarie University, Australia, and Professor of Applied Linguistics at the Open University, UK. At Macquarie, he has been Chair of the Department of Linguistics; he established and was Executive Director of the National Centre for English Language Teaching and Research (NCELTR) and foundational Director of the Centre for Language in Social Life (CLSL). He has written or edited over 150 publications and co-edits the *Journal of Applied Linguistics*. From 1996 to 2002 he was President of the International Association of Applied Linguistics (AILA). He has acted as a consultant in more than thirty-five countries and as external faculty assessor in thirty-six universities worldwide.

Ronald Carter is Professor of Modern English Language in the School of English Studies at the University of Nottingham. He has published extensively in applied linguistics, literary studies and language in education, and has written or edited over forty books and a hundred articles in these fields. He has given consultancies in the field of English language education, mainly in conjunction with the British Council, in over thirty countries worldwide, and is editor of the Routledge Interface series and advisory editor to the Routledge English Language Introduction series. He was recently elected a fellow of the British Academy of Social Sciences and is currently UK Government Advisor for ESOL and Chair of the British Association of Applied Linguistics (BAAL).

TITLES IN THE SERIES

Intercultural Communication: An advanced resource book
Adrian Holliday, Martin Hyde and John Kullman

Translation: An advanced resource book
Basil Hatim and Jeremy Munday

Grammar and Context: An advanced resource book
Ann Hewings and Martin Hewings

Second Language Acquisition: An advanced resource book
Kees de Bot, Wander Lowie and Marjolijn Verspoor

Corpus-based Language Studies: An advanced resource book
Anthony McEnery, Richard Xiao and Yukio Tono

Language and Gender: An advanced resource book
Jane Sunderland

English for Academic Purposes: An advanced resource book
Ken Hyland

Language Testing and Assessment: An advanced resource book
Glenn Fulcher and Fred Davidson

Language Testing and Assessment

An advanced resource book

Glenn Fulcher and Fred Davidson

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2007
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Simultaneously published in the USA and Canada
by Routledge
270 Madison Ave, New York, NY 10016

Routledge is an imprint of the Taylor & Francis Group, an informa business

This edition published in the Taylor & Francis e-Library, 2006.

"To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk."

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
Fulcher, Glenn.
Language testing and assessment / Glenn Fulcher & Fred Davidson.
p. cm.
Includes bibliographical references and index.
1. Language and languages—Ability testing. I. Davidson, Fred. II. Title.
P53.4.F85 2007
418.0076—dc22 2006022928

ISBN 0-203-44906-1 Master e-book ISBN

ISBN10: 0–415–33946–4 (hbk)
ISBN10: 0–415–33947–2 (pbk)
ISBN10: 0–203–44906–1 (ebk)

ISBN13: 978–0–415–33946–9 (hbk)
ISBN13: 978–0–415–33947–6 (pbk)
ISBN13: 978–0–203–44906–6 (ebk)

For Jenny and Robin

Contents

List of figures and tables xiv
Series editors' preface xv
Acknowledgments xvii
How to use this book xix

SECTION A: INTRODUCTION **1**

Unit A1 Introducing validity **3**
A1.1 Introduction 3
A1.2 Three 'types' of validity in early theory 4
A1.3 Cutting the validity cake 12
Summary 21

Unit A2 Classroom assessment **23**
A2.1 Introduction 23
A2.2 Pedagogy and the measurement paradigm 25
Summary 35

Unit A3 Constructs and models **36**
A3.1 Introduction 36
A3.2 The nature of models 37
A3.3 Canale and Swain's model of communicative competence 38
A3.4 Canale's adaptations 39
A3.5 Bachman's model of communicative language ability (CLA) 42
A3.6 Celce-Murcia, Dörnyei and Thurrell's model of communicative competence 47
A3.7 Interactional competence 49
A3.8 From models to frameworks: validity models and performance conditions 50
Summary 51

Unit A4 Test specifications and designs **52**
A4.1 Introduction 52
A4.2 Planning in test authoring 53
A4.3 Guiding language versus samples 54
A4.4 Congruence (or fit-to-spec) 55
A4.5 How do test questions originate? Reverse engineering and archetypes 56
A4.6 Reverse engineering 57
A4.7 Where do test items come from? What is the true genesis of a test question? 58
A4.8 Spec-driven test assembly, operation and maintenance 59

A4.9 Towards spec-driven theory 60
Summary 61

Unit A5 Writing items and tasks 62
A5.1 Introduction 62
A5.2 Evidence-centred design (ECD) 63
A5.3 Describing items and tasks 69
A5.4 Tasks and teaching 73
Summary 75

Unit A6 Prototypes, prototyping and field tests 76
A6.1 Introduction 76
A6.2 Prototypes 76
A6.3 Prototyping 79
A6.4 Field testing 85
A6.5 The iterative nature of the process 89
Summary 89

Unit A7 Scoring language tests and assessments 91
A7.1 Introduction 91
A7.2 Defining the quality of language 93
A7.3 Developing scoring systems 96
A7.4 Intuition and data 98
A7.5 Problems with scales 98
A7.6 Scoring in classical test theory 101
A7.7 Reliability 104
A7.8 Score transformations 108
A7.9 Item response theory 109
A7.10 Endowing a score with special meaning 111
Summary 114

Unit A8 Administration and training 115
A8.1 Introduction 115
A8.2 Getting things done 117
A8.3 Quality management systems 127
A8.4 Constraints 128
A8.5 Test administration within the ECD delivery model 129
A8.6 Rater and interlocutor training 131
A8.7 Security 132
A8.8 Test administration for disabled people 135
Summary 137

Unit A9 Fairness, ethics and standards 138
A9.1 Introduction 138
A9.2 Professionalism as a community of practitioners 138
A9.3 Professionalism and democracy 141
A9.4 Consequentialism 142
A9.5 On power and pessimism 144
A9.6 Professional conduct: standards for practice 155
A9.7 Responsibilities of language testers and their limitations 156
A9.8 Accountability 157
Summary 158

Unit A10 Arguments and evidence in test validation and use 159
A10.1 Introduction 159
A10.2 Argumentation as solution 162
A10.3 The form of an argument 164
A10.4 Argument in evidence-centred design 167
A10.5 Arguments in language testing 168
A10.6 Arguments and feasibility 176
A10.7 Argument, evidence and ethics 176
Summary 178

SECTION B: EXTENSION 179

Unit B1 Construct validity 181
Cronbach, L. J. and Meehl, P. E. 'Construct validity in psychological tests' 182

Unit B2 Pedagogic assessment 192
Moss, P. 'Reconceptualizing validity for classroom assessment' 193

Unit B3 Investigating communicative competence 203
Canale, M. and Swain, M. 'Theoretical bases of communicative approaches to second language teaching and testing' 203

Unit B4 Optimal specification design 212
Davidson, F. and Lynch, B. K. *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications* 212

Unit B5 Washback 221
Alderson, J. C. and Wall, D. 'Does washback exist?' 222

Unit B6 Researching prototype tasks 230
Cumming, A., Grant, L., Mulcahy-Ernt, P. and Powers, D. *A Teacher-Verification Study of Speaking and Writing Prototype Tasks for a New TOEFL* 230

Unit B7 Scoring performance tests 249
Hamp-Lyons, L. 'Scoring procedures for ESL contexts' 250

Unit B8 Interlocutor training and behaviour 258
Brown, A. 'Interviewer variation and the co-construction of speaking proficiency' 260

Unit B9 Ethics and professionalism 270
Davies, A. 'Demands of being professional in language testing' 270

Unit B10 Validity as argument 278
Kane, M. T. (1992). 'An argument-based approach to validity' 278

SECTION C: EXPLORATION **291**

Unit C1 Validity – an exploration 293
Unit C2 Assessment in school systems 298
Unit C3 What do items really test? 304
Unit C4 Evolution in action 312
Unit C5 To see a test in a grain of sand . . . 320
Unit C6 Analysing items and tasks 326
Unit C7 Designing an alternative matrix 333
Unit C8 Administration and alignment 343
Unit C9 In a time far, far away . . . 352
Unit C10 To boldly go 361

Glossary 369
Notes 378
References 379
Index 396

Contents cross-referenced

Section A: Introduction

Unit A1 Introducing validity 3
A1.1 Introduction 3
A1.2 Three ‘types’ of validity in early theory 4
A1.3 Cutting the validity cake 12
Summary 21

Unit A2 Classroom assessment 23
A2.1 Introduction 23
A2.2 Pedagogy and the measurement paradigm 25
Summary 35

Unit A3 Constructs and models 36
A3.1 Introduction 36
A3.2 The nature of models 37
A3.3 Canale and Swain’s model of communicative competence 38
A3.4 Canale’s adaptations 39
A3.5 Bachman’s model of communicative language ability (CLA) 42
A3.6 Celce-Murcia, Dörnyei and Thurrell’s model of communicative competence 47
A3.7 Interactional competence 49
A3.8 From models to frameworks: validity models and performance conditions 50
Summary 51

Unit A4 Test specifications and designs 52
A4.1 Introduction 52
A4.2 Planning in test authoring 53
A4.3 Guiding language versus samples 54
A4.4 Congruence (or fit-to-spec) 55
A4.5 How do test questions originate? Reverse engineering and archetypes 56
A4.6 Reverse engineering 57
A4.7 Where do test items come from? What is the true genesis of a test question? 58
A4.8 Spec-driven test assembly, operation and maintenance 59
A4.9 Towards spec-driven theory 60
Summary 61

Unit A5 Writing items and tasks 62
A5.1 Introduction 62
A5.2 Evidence-centred design (ECD) 63
A5.3 Describing items and tasks 69
A5.4 Tasks and teaching 73
Summary 75

Unit A6 Prototypes, prototyping and field tests 76
A6.1 Introduction 76
A6.2 Prototypes 76
A6.3 Prototyping 79
A6.4 Field testing 85
A6.5 The iterative nature of the process 89
Summary 89

Unit A7 Scoring language tests and assessments 91
A7.1 Introduction 91
A7.2 Defining the quality of language 93
A7.3 Developing scoring systems 96
A7.4 Intuition and data 98
A7.5 Problems with scales 98
A7.6 Scoring in classical test theory 101
A7.7 Reliability 104
A7.8 Score transformations 108
A7.9 Item response theory 109
A7.10 Endowing a score with special meaning 111
Summary 114

Unit A8 Administration and training 115
A8.1 Introduction 115
A8.2 Getting things done 117
A8.3 Quality management systems 127
A8.4 Constraints 128
A8.5 Test administration within the ECD delivery model 129
A8.6 Rater and interlocutor training 131
A8.7 Security 132
A8.8 Test administration for disabled people 135
Summary 137

Unit A9 Fairness, ethics and standards 138
A9.1 Introduction 138
A9.2 Professionalism as a community of practitioners 138
A9.3 Professionalism and democracy 141
A9.4 Consequentialism 142
A9.5 On power and pessimism 144
A9.6 Professional conduct: standards for practice 155
A9.7 Responsibilities of language testers and their limitations 156
A9.8 Accountability 157
Summary 158

Unit A10 Arguments and evidence in test validation and use 159
A10.1 Introduction 159
A10.2 Argumentation as solution 162
A10.3 The form of an argument 164
A10.4 Argument in evidence-centred design 167
A10.5 Arguments in language testing 168
A10.6 Arguments and feasibility 176
A10.7 Argument, evidence and ethics 176
Summary 178

Section B: Extension

Unit B1 **Construct validity** **181**
Cronbach, L. J. and Meehl, P. E. 'Construct validity in psychological tests' 182

Unit B2 **Pedagogic assessment** **192**
Moss, P. 'Reconceptualizing validity for classroom assessment' 193

Unit B3 **Investigating communicative competence** **203**
Canale, M. and Swain, M. 'Theoretical bases of communicative approaches to second language teaching and testing' 203

Unit B4 **Optimal specification design** **212**
Davidson, F. and Lynch, B. K. *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications* 212

Unit B5 **Washback** **221**
Alderson, J. C. and Wall, D. 'Does Washback Exist?' 222

Unit B6 **Researching prototype tasks** **230**
Cumming, A., Grant, L., Mulcahy-Ernt, P. and Powers, D. *A Teacher-Verification Study of Speaking and Writing Prototype Tasks for a New TOEFL* 230

Unit B7 **Scoring performance tests** **249**
Hamp-Lyons, L. 'Scoring procedures for ESL contexts' 250

Unit B8 **Interlocutor training and behaviour** **258**
Brown, A. 'Interviewer variation and the co-construction of speaking proficiency' 260

Unit B9 **Ethics and professionalism** **270**
Davies, A. 'Demands of being professional in language testing' 270

Unit B10 **Validity as argument** **278**
Kane, M. T. (1992). 'An argument-based approach to validity' 278

Section C: Exploration

Unit C1 **Validity – an exploration** **293**

Unit C2 **Assessment in school systems** **298**

Unit C3 **What do items really test?** **304**

Unit C4 **Evolution in action** **312**

Unit C5 **To see a test in a grain of sand . . .** **320**

Unit C6 **Analysing items and tasks** **326**

Unit C7 **Designing an alternative matrix** **333**

Unit C8 **Administration and alignment** **343**

Unit C9 **In a time far, far away . . .** **352**

Unit C10 **To boldly go** **361**

Figures and tables

FIGURES

A3.1 Models, frameworks and specifications 37
A3.2 Canale's adaptation of the Canale and Swain model 41
A3.3 Components of communicative language ability in language use 42
A3.4 Components of language competence 43
A3.5 Some components of language use and language test performance 46
A3.6 The Celce-Murcia et al. model of communicative competence 47
A5.1 Models in the conceptual assessment framework of ECD 68
A7.1 The *what* and *how much/how good* of language testing 91
A7.2 The structure of the CEF 99
A7.3 Three possible cut scores 113
A8.1 Four-process delivery model 130
A10.1 The basic form of an argument 165
A10.2 The argument of Harry's nationality 166
A10.3 Mislevy's language testing example 167
A10.4 An argument structure for textual competence items 169
A10.5 The cloze argument 170
A10.6 A test use argument structure 173

TABLES

A1.1 Facets of validity 13
A6.1 Basic item statistics for five items included on a prototype test of textual competence 87
A7.1 Responses of twenty test takers to ten test items 101
A7.2 Item facility values 102
A7.3 Point biserial correlations 104
A7.4 Rasch difficulty estimates for ten items 110
A7.5 Ability estimates for twenty test takers 110
C6.1 Item statistics 326
C7.1 Ethelynn's term-final results 334
C7.2 Data from the alternative universe 336
C7.3 Data mining structure 339

Series editors' preface

This series provides a comprehensive guide to a number of key areas in the field of applied linguistics. Applied linguistics is a rich, vibrant, diverse and essentially interdisciplinary field. It is now more important than ever that books in the field provide up-to-date maps of ever-changing territory.

The books in this series are designed to give key insights into core areas. The design of the books ensures, through key readings, that the history and development of a subject is recognized while, through key questions and tasks, integrating understandings of the topics, concepts and practices that make up its essentially interdisciplinary fabric. The pedagogic structure of each book ensures that readers are given opportunities to think, discuss, engage in tasks, draw on their own experience, reflect, research and to read and critically re-read key documents.

Each book has three main sections, each made up of approximately ten units:

A: An **Introduction** section: in which the key terms and concepts are introduced, including introductory activities and reflective tasks, designed to establish key understandings, terminology, techniques of analysis and the skills appropriate to the theme and the discipline.

B: An **Extension** section: in which selected core readings are introduced (usually edited from the original) from existing books and articles, together with annotations and commentary, where appropriate. Each reading is introduced, annotated and commented on in the context of the whole book, and research/follow-up questions and tasks are added to enable fuller understanding of both theory and practice. In some cases, readings are short and synoptic and incorporated within a more general exposition.

C: An **Exploration** section: in which further samples and illustrative materials are provided with an emphasis, where appropriate, on more open-ended, student-centred activities and tasks, designed to support readers and users in undertaking their own locally relevant research projects. Tasks are designed for work in groups or for individuals working on their own.

The books also contain a glossary or glossarial index and a detailed, thematically organized A–Z guide to the main terms used in the book, which lays the ground for further work in the discipline. There are also annotated guides to further reading and extensive bibliographies.

The target audience for the series is upper undergraduates and postgraduates on language, applied linguistics and communication studies programmes as well as teachers and researchers in professional development and distance learning programmes. High-quality applied research resources are also much needed for teachers of EFL/ESL and foreign language students at higher education colleges and universities worldwide. The books in the Routledge Applied Linguistics series are aimed at the individual reader, the student in a group and at teachers building courses and seminar programmes.

We hope that the books in this series meet these needs and continue to provide support over many years.

The Editors

Professor Christopher N. Candlin and Professor Ronald Carter are the series editors. Both have extensive experience of publishing titles in the fields relevant to this series. Between them they have written and edited over one hundred books and two hundred academic papers in the broad field of applied linguistics. Chris Candlin was president of AILA (International Association for Applied Linguistics) from 1997 to 2003 and Ron Carter is Chair of BAAL (British Association for Applied Linguistics) from 2003 to 2006.

Professor Christopher N. Candlin
Senior Research Professor
Department of Linguistics
Division of Linguistics and Psychology
Macquarie University
Sydney NSW 2109
Australia

and

Professor of Applied Linguistics
Faculty of Education and Language Studies
The Open University
Walton Hall
Milton Keynes MK7 6AA
UK

Professor Ronald Carter
School of English Studies
University of Nottingham
Nottingham NG7 2RD
UK

Acknowledgments

We would like to thank the series editors, Chris Candlin and Ron Carter, for their timely feedback on drafts of this manuscript, and their constant encouragement and help. At every stage in the writing process their wisdom and insight have been an inspiration, and we could not have asked for better editors with whom to work. At Routledge we would also like to thank Louisa Semlyen and Nadia Seemungal for their help, advice and efficiency.

The picture of the Golden Gate Bridge in Unit A4 is taken from http://www.sonic.net/~playland/ggsag.html. This picture is by an unknown photographer and is copyright-free, but we wish to acknowledge the source.

The examples of rapid prototyping in manufacturing in Unit A6 are reproduced by kind permission of Laser Prototypes Europe Ltd. The authors wish to thank Mr Tom Walls for his cooperation and advice while writing this part of Unit A6 (Tom@laserproto.com).

We would like to thank Mr R. J. Hann, erstwhile Staff Sergeant of Delta Company Fourth Battalion (4 RAR/NZ (ANZAC)), for permission to publish the picture in Unit C3, taken circa 1970 and published at http://www.diggerhistory.info/pages-uniforms/slouch_hat-b.htm.

The PERT and Gantt Charts used in Unit C6 are reproduced with permission of Alan Levine of Maricopa Community College.

The extracts in Section B are reproduced by kind permission of the publishers, as follows.

Unit B2
Moss, P. A. (2003) 'Reconceptualizing validity for classroom assessment.' *Educational Measurement: Issues and Practice* 22, 4, 13–25. Reproduced by permission of Blackwell Publishing.

Unit B3
Canale, M. and Swain, M. (1980) 'Theoretical bases of communicative approaches to second language teaching and testing.' *Applied Linguistics* 1, 1, 1–47 by permission of Oxford University Press.

Unit B4
Davidson, F. and Lynch, B. K. (2002) Chapter 3: 'Problems and issues in specification writing.' In *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications.* Copyright © 2002 by Yale University. Reproduced by permission of Yale University Press.

Unit B5
Alderson, J. C. and Wall, D. (1993) 'Does washback exist?' *Applied Linguistics* 14, 2, 115–129 by permission of Oxford University Press.

Unit B6
TOEFL materials selected from TOEFL Monograph Series MS-26 – *A Teacher-Verification Study of Speaking and Writing Prototype Tasks for a New TOEFL*, Educational Testing Service 2005. Reprinted by permission of Educational Testing Service, the copyright owner. However, the test questions and any other testing information are provided in their entirety by Routledge. No endorsement of this publication by Educational Testing Service should be inferred.

Unit B7
Hamp-Lyons, L. (1993) 'Scoring procedures for ESL contexts.' In Hamp-Lyons, L. (ed.) *Assessing Second Language writing in Academic Contexts.* Norwood, NJ: Ablex, 241–276. Copyright © 1991 by Ablex Publishing Corporation. Reproduced with permission of Greenwood Publishing Group, Inc., Westport, CT.

Unit B8
Reproduced with permission from Brown, A. (2003) 'Interviewer variation and the co-construction of speaking proficiency.' *Language Testing* 20, 1, 1–25. Copyright © Sage Publications 2003, by permission of Sage Publications Ltd and the author.

Unit B9
Reproduced with permission from Davies, A. (1997) 'Demands of being professional in language testing.' *Language Testing* 14, 3, 328–339. Copyright © Sage Publications Ltd 1997, by permission of Sage Publications Ltd.

Unit B10
Kane, M. T. (1992) 'An argument-based approach to validity.' *Psychological Bulletin* 112, 527–535. Copyright © 1992 by the American Psychological Association. Reprinted with permission.

How to use this book

Testing and assessment are part of modern life. Schoolchildren around the world are constantly assessed, whether to monitor their educational progress, or for governments to evaluate the quality of school systems. Adults are tested to see if they are suitable for a job they have applied for, or if they have the skills necessary for promotion. Entrance to educational establishments, to professions and even to entire countries is sometimes controlled by tests. Tests play a fundamental and controversial role in allowing access to the limited resources and opportunities that our world provides. The importance of understanding *what* we test, *how* we test and the *impact* that the use of tests has on individuals and societies cannot be overstated. Testing is more than a technical activity; it is also an ethical enterprise.

The practice of language testing draws upon, and also contributes to, all disciplines within applied linguistics. However, there is something fundamentally different about language testing. Language testing is all about building better tests, researching how to build better tests and, in so doing, understanding better the things that we test.

Sociolinguists do not create 'sociolinguistic things'. Discourse analysts do not create discourses. Phonologists do not create spoken utterances. Language testing, in contrast, is about *doing*. It is about *creating tests*.

In a sense, therefore, each section of this book is about the practical aspects of *doing* and of *creating*. And so each section has a research implication; no section is concerned purely with exposition. Research ideas may be made explicit in the third section, *Exploration*, but they are implicit throughout the book; put another way, the creative drive of language testing makes it a research enterprise, we think, at all times.

In the text we do not merely reflect the state of the art in language testing and assessment; nor do we simply introduce existing research. Our discussion is set within a new approach that we believe brings together testing practice, theory, ethics and philosophy. At the heart of our new approach is the concept of *effect-driven testing*. This is a view of test validity that is highly pragmatic. Our emphasis is on the outcome of testing activities. Our concern with test effect informs the order and structure of chapters, and it defines our approach to test design and development.

As test design and development is about *doing*, *creating* and *researching*, we have taken special care over the activities. With Dewey, we believe that through *doing* we grow as language testers, as applied linguists and as language teachers.

The book is divided into three sections. *A: Introduction* consists of ten units dealing with the central concepts of language testing and assessment. It contains activities for you to carry out alone, or with others if you are studying this book as part of a course. *B: Extension* provides extracts from articles or books relating to language testing and assessment which give you further insights into the concepts introduced in Section A. Each extract in *B: Extension* is accompanied by activities to focus your reading and help you to evaluate critically what you have read and understand how it links to a wider discussion of language testing and assessment. *C: Exploration* builds on the material you will already have found in the book. In this section we provide extended activities that help you to work through practical and theoretical problems that have been posed in the other sections. We also present ideas for individual and group project work, as well as suggestions for research projects.

The organization of this book allows you to concentrate on particular themes, such as *classroom assessment* or *writing items and tasks*, by reading the relevant units from *A: Introduction*, *B: Extension* and *C: Exploration* consecutively. Alternatively, you may wish to read the whole of *A: Introduction* before embarking on Sections B and C. In fact, you may decide to read the Sections in any sequence, just as you would read Julio Cortázar's novel *Hopscotch*: there is no one right place to start, and each path through the text provides a different experience. Whichever choice you make, the book is extensively cross-referenced and carefully indexed so that you can easily find your way around the material.

At the end of the book we provide a glossary of key terms that are not explained within the text itself. If you come across a term about which you feel uncertain, simply turn to the glossary for an explanation. We also provide an extensive list of references for additional reading.

In addition to the book itself, there is also a website http://www.routledge.com/textbooks/9780415339476 in which we provide very extensive additional reading, activities, links to relevant websites and further ideas for projects that you might like to undertake on your own or with colleagues.

SECTION A
Introduction

Unit A1
Introducing validity

A1.1 INTRODUCTION

Every book and article on language testing deals to some extent with validity. It is the central concept in testing and assessment, and so comes at the very beginning of this book. In other texts, it normally appears anywhere from chapter 4 to chapter 8. But this positioning implies that validity enquiry is something that is 'done' after a test or assessment has been written and is in use. This is to misunderstand the importance of validity. In this first chapter we are going to investigate the concept of validity. We are not going to shy away from asking serious questions about what it means, and why it is important. Only through tackling the most difficult topic first does everything else fall into place so much more easily.

Questions of validity impact on our daily lives and how we interact with people and the world around us; it is just that we don't reflect very frequently on the kinds of validity decisions that we make. We observe all kinds of behaviour, hear what people say to us and make inferences that lead to action or beliefs. One of the most pressing validity issues for humans is 'Does s/he love me?' The concept of 'love' is one that is virtually impossible to define, which is why it generates so much poetry and nearly every song ever written. The validity question a person faces when asking this question is: on the basis of what this person says and does, can I infer a set of feelings and attitudes that will justify me in taking decisions which, if I get it wrong, could lead to unwanted (and potentially disastrous) consequences?

But in our everyday lives we don't put validity questions formally, or try to list the kinds of evidence that we would need to collect before falling in love! In language testing this is precisely what we have to do, so that we can produce a chain of reasoning and evidence from what we think a test score means, and the actions we intend to take on the basis of that inference, back to the skills, abilities or knowledge that any given test taker may have. The closest we have to this for love is possibly the work of Stendhal (1975), who notes that in the infancy of love

> The lover's mind vacillates between three ideas:
>
> 1 She is perfect.
> 2 She loves me.
> 3 How can I get the strongest possible proofs of her love?

He goes on to explore the ways in which humans gather the evidence they need to ‘dispel doubt’. In language testing this dispelling of doubt is removing as much uncertainty as possible that the scores mean what we think they mean, so that we can take actions without the fear of making serious mistakes. It is deliberate and planned, while in love, as other areas of life, it is intuitive and most often unconscious.

‘Validity’ in testing and assessment has traditionally been understood to mean discovering whether a test ‘measures accurately what it is intended to measure’ (Hughes, 1989: 22), or uncovering the ‘appropriateness of a given test or any of its component parts as a measure of what it is purposed to measure’ (Henning, 1987: 170). This view of validity presupposes that when we write a test we have an *intention* to measure something, that the ‘something’ is ‘real’, and that validity enquiry concerns finding out whether a test ‘actually does measure’ what is intended. These are assumptions that were built into the language of validity studies from the early days, but ones that we are going to question.

In this Unit we will take a historical approach, starting with early validity theory that was emerging after the Second World War, and trace the changes that have occurred since then. We will attempt to explain the terminology, and provide examples that will help to make the subject look a little less daunting than is usually the case.

A1.2 THREE ‘TYPES’ OF VALIDITY IN EARLY THEORY

In the early days of validity investigation, validity was broken down into three ‘types’ that were typically seen as distinct. Each type of validity was related to the kind of evidence that would count towards demonstrating that a test was valid. Cronbach and Meehl (1955) described these as:

- Criterion-oriented validity
  - Predictive validity
  - Concurrent validity
- Content validity
- Construct validity

We will introduce each of these in turn, and then show how this early approach has changed.

A1.2.1 Criterion-oriented validity

When considering criterion-oriented validity, the tester is interested in the relationship between a particular test and a criterion to which we wish to make predictions. For example, I may wish to predict from scores on a test of second-language academic reading ability whether individuals can cope with first-semester undergraduate business studies texts in an English-medium university. What we are really interested in here is the criterion, whatever it is that we wish to know about, but for which we don't have any direct evidence. In the example above we cannot see whether future students can do the reading that will be expected of them before they actually arrive at the university and start their course.

In this case the validity evidence is the strength of the predictive relationship between the test score and that performance on the criterion. Of course, it is necessary to decide what would count as 'ability to cope with' – as it is something that must be measurable. Defining precisely what we mean by such words and phrases is a central part of investigating validity.

Task A1.1

Consider the following situations where you may wish to use a test to discover something about your students:

How many students in my class are likely to pass the Certificate of Proficiency at the end of the semester?

If Mr Hassan starts work as an air traffic controller now, will he be able to successfully guide aircraft out of danger in near-miss situations?

My students of legal English are going to go on work experience later in the year. How do I know whether they will be able to help prepare the paperwork for court cases?

I need to plan next semester's syllabus for my class. I need to discover which elements of this semester's syllabus I need to recycle.

➤ In each case what would you use as a criterion (or criteria), and why?

➤ Try to think of other examples from your own teaching situation.

Predictive validity is the term used when the test scores are used to predict some future criterion, such as academic success. If the scores are used to predict a criterion at the same time the test is given, we are studying *concurrent validity*.

Returning to the example given above, let us assume that in this case 'ability to cope' is defined as a subject tutor's judgment of whether students can adequately read set texts to understand lectures and write assignments. We might be interested in discovering the relationship between students' scores on our test prior to starting academic studies and the judgments of the tutors once the students have started their programme. This would be a *predictive validity study*. We would hope that we could identify a score on the reading test above which tutors would judge readers to be competent, and below which they would judge some readers to lack the necessary reading skills for academic study. This would be the 'cut score' for making a predictive decision about the likelihood of future success on the criterion.

Suppose that my reading test is too long, and for practical purposes it needs to be made much shorter. As we know that shorter tests mean that we collect less evidence about reading ability, one of the questions we would wish to ask is to what extent the shorter test is capable of predicting the scores on the longer test. In other words, could the shorter test replace the larger test and still be useful? This would be an example of a *concurrent validity study* that uses the longer test as the criterion.

A1.2.2 Content validity

Content validity is defined as any attempt to show that the content of the test is a representative sample from the domain that is to be tested. In our example of the academic reading test it would be necessary to show that the texts selected for the test are typical of the types of texts that would be used in first-year undergraduate business courses. This is usually done using expert judges. These may be subject teachers, or language teachers who have many years' experience in teaching business English. The judges are asked to look at texts that have been selected for inclusion on the test and evaluate them for their representativeness within the content area. Secondly, the items used on the test should result in responses to the text from which we can make inferences about the test takers' ability to process the texts in ways expected of students on their academic courses. For example, we may discover that business students are primarily required to read texts to extract key factual information, take notes and use the notes in writing assignments. In our reading test we would then try to develop items that tap the ability to identify key facts.

Carroll (1980: 67) argued that achieving content validity in testing English for Academic Purposes (EAP) consisted of describing the test takers, analysing their 'communicative needs' and specifying test content on the basis of their needs. In early approaches to communicative language testing the central issue in establishing content validity was how best to 'sample' from needs and the target domain (Fulcher, 1999a: 222–223).

Task A1.2

➤ Consider these target domains. For each, try to list what a test may need to contain to be relevant to that domain.

1 Nursing in a hospital
2 Staffing the reception in a hotel
3 Check-in desk at an international airport
4 Taxi driver in a capital city
5 Tour guide in a tourist resort.

- ➤ Do you have students for whom the content domain can easily be defined?
- ➤ What makes it very difficult to define a content domain?

A1.2.3 Construct validity

The first problem with construct validity is defining what a 'construct' is. Perhaps the easiest way to understand the term 'construct' is to think of the many abstract nouns that we use on a daily basis, but for which it would be extremely hard to point to an example. Consider these, the first of which we have already touched on.

1. Love
2. Intelligence
3. Anxiety
4. Thoughtfulness
5. Fluency
6. Aptitude
7. Extroversion
8. Timidity
9. Persuasiveness
10. Empathy.

As we use these terms in everyday life we have no need to define them. We all assume that we know what they mean, and that the meaning is shared. So we can talk with our friends about how much empathy someone we know may have, or how fluent a speaker someone is. But this is to talk at the level of everyday concepts. For a general term to become a construct, it must have two further properties. Firstly, it must be defined in such a way that it becomes measurable. In order to measure 'fluency' we have to state what we could possibly observe in speech to make a decision about whether a speaker is fluent. It turns out that many people have different definitions of fluency, ranging from simple speed of speech, to lack of hesitation (or strictly 'pauses', because 'hesitation' is a construct itself), to specific observable features of speech (see Fulcher, 1996). Secondly, any construct should be defined in such a way that it can have relationships with other constructs that are different. For example, if I generate descriptions of 'fluency' and 'anxiety' I may hypothesize that, as anxiety increases, fluency will decrease, and vice versa. If this hypothesis is tested and can be supported, we have the very primitive beginnings of a theory of speaking that relates how we perform to emotional states.

To put this another way, concepts become constructs when they are so defined that they can become 'operational' – we can measure them in a test of some kind by linking the term to something observable (whether this is ticking a box or performing some communicative action), and we can establish the place of a construct in a theory that relates one construct to another (Kerlinger and Lee, 2000: 40), as in the case of fluency and anxiety above.

A1.2.4 Construct validity and truth

In the early history of validity theory there was an assumption that there is such a thing as a 'psychologically real construct' that has an independent existence in the test taker, and that the test scores represent the degree of presence or absence of this very real property. As Cronbach and Meehl (1955: 284) put it:

> Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim.

This brings us to our first philosophical observation. It has frequently been argued that early validity theorists were positivistic in their outlook. That is, they assumed that their constructs actually existed in the heads of the test takers. Again, Cronbach and Meehl (1955: 284) state: 'Scientifically speaking, to "make clear what something is" means to set forth the laws in which it occurs. We shall refer to the interlocking system of laws which constitute a theory as a nomological network.'

The idea of a nomological network is not difficult to grasp. Firstly, it contains a number of constructs, and their names are abstract, like those in the list above. In language teaching and testing, 'fluency' and 'accuracy' are two well-known constructs. Secondly, the nomological network contains the observable variables – those things that we can see and measure directly, whereas we cannot see 'fluency' and 'accuracy' directly.

What might these observable variables be? Whatever we choose makes up the definition of the constructs. For fluency we may wish to observe speed of delivery or the number of unfilled pauses, for example. For accuracy, we could look at the ratio of correct to incorrect tense use or word order. From what we can observe, we then make an inference about how 'fluent' or how 'accurate' a student's use of the second language is.

The network is created by asking what we expect the relationship between 'fluency' and 'accuracy' to be. One hypothesis could be that in speech, as fluency increases, accuracy decreases, because learners cannot pay attention to form when the demands of processing take up all the capacity of short-term memory. Another hypothesis could be that, as accuracy increases, the learner becomes more fluent, because language form has become automatic. Stating this kind of relationship between constructs therefore constitutes a theory, and theory is very powerful. Even in this simple example we could now set out a testable research hypothesis: fluency and accuracy are inversely related in students below X level of proficiency, and above it they are positively related.

Let us see if we can relate this back to our example from everyday life.

Task A1.3

Here we will set out Stendhal's theory of love as if it were a nomological network. Constructs:

1 Passionate Love, 'like that of Heloïse for Abelard'
2 Mannered Love, 'where there is no place for anything at all unpleasant – for that would be a breach of etiquette, of good taste, of delicacy, and so forth'
3 Physical Love, 'where your love life begins at sixteen'
4 Vanity Love, in which 'men . . . both desire and possess a fashionable woman, much in the way one might own a fine horse'.

➤ What do you think are the possible relationships between these four constructs?

For example, assuming that I could measure these types of love, I might hypothesize that as the strength of mannered love increases, passionate love might decrease. I may further hypothesize that there is a strong positive relationship between physical love and passionate love, and only a weak relationship between mannered love and physical love.

➤ Write down a number of hypotheses.

Stendhal went on to attach certain observable behaviours to each 'type' of love. Here are some of them. Which of these observable behaviours do you think Stendhal thought characterized each type of love?

- Behaviour always predictable
- Lack of concentration
- Always trying to be witty in public
- Staring at girls
- Following habits and routines carefully
- Always very money-conscious
- Engaging in acts of cruelty
- Touching.

➤ Try to list other behaviours that may be typical of a type of love as described by Stendhal.

Is your nomological net a satisfying theory of love? Probably not. Stendhal himself wrote: 'Instead of defining four kinds of love, one might well admit eight or ten distinctions. There are perhaps as many different ways of feeling as there are of seeing.'

➤ What are the implications of this for construct definition in language testing?

In philosophy, the logical positivists (some of whom Cronbach and Meehl reference) argued that only propositions that could be verified relative to empirical evidence were meaningful, and that all other propositions were not just false but actually meaningless (Ayer, 1936). In our examples of nomological networks above, meaning is created by measuring the variables (unfilled pauses, or predictability of actions, for example) and testing how these relate to the constructs that they define in terms of a theory that establishes relationships among constructs.

In testing and assessment this meant that if there is no possible way to test the hypotheses created by the relationship between observable variables, observable variables and constructs, and between constructs, the theory is meaningless, or not 'scientifically admissible'.

The underlying philosophical assumptions have been heavily criticized, and in 1989 Cronbach himself said that the position of 1955 was 'pretentious'. However, there were elements in the 1955 work that have continued to influence validity research – particularly the argument that construct definition lies at the centre of testing and assessment, and that at the heart of any validity study is the investigation of the intended meaning and interpretation of test scores. And central to understanding score meaning lies the question of what evidence can be presented to support a particular score interpretation. There is also one other aspect of the 1955 work that is still important. Cronbach and Meehl argue that it is necessary to institute a programme of research to collect the evidence that will be used to support specific interpretations, and 'make the evidence for the claim public' so that it can be evaluated by the community of researchers. They argue that 'confidence in a theory is increased as more relevant evidence confirms it, but it is always possible that tomorrow's investigation will render the theory obsolete'.

This is not positivistic in tone. It recognizes that our present knowledge and theories are tenuous and temporal, even incorrect. But they represent our 'best shot' at understanding what we wish to test, given our imperfect theories. The notion of the nomological network and the testability of hypotheses between variables and constructs to form theories was an early attempt to ensure that theory building was driven by data and the 'scientific method'.

This takes us to the heart of epistemology and what it means to say that something is 'true' or 'real'. In 1877 C. S. Peirce had put forward a pragmatic notion of meaning: 'Consider what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object' (Peirce, 1877: 146). To translate this into modern English: if we believe something to be true, will the effect be that we are better able to understand the world around us and use the idea to do something practical in a way that results in progress? Or as Messick (1989: 26) puts it (using the term 'instrumentalist' for 'pragmatist'): 'According to the instrumentalist theory of truth, a statement is true if it is useful in directing inquiry or guiding action.' Cronbach and Meehl could easily have argued that if a nomological network allows us to make better decisions in testing and assessment, then it is 'contingently true' (because it is practically useful) until it is shown to be a partial or inadequate explanation. The alternatives, Peirce argues, are believing something to be true on other spurious grounds, such as 'that's the way it's always been', or because the person who puts forward the theory is the most authoritative in the field at the moment. Messick (1989: 23) also added from a post-positivistic era that:

> Nomological networks are viewed as an illuminating way of speaking systematically about the role of constructs in psychological theory and measurement, but not as the only way. The nomological framework offers a useful guide for disciplined thinking about the process of validation but cannot serve as the prescriptive validation model to the exclusion of other approaches.

This quotation shows another shift in thinking about validation. The nomological network is just one approach to addressing validity questions. It is just one of the tools at our disposal, but there are many others that would yield validity evidence.

Secondly, Peirce held that theories may evolve or be overthrown by a community of researchers, and that with passing time, theories will evolve and become more adequate in their usability:

> This great law is embodied in the conception of truth and reality. The opinion that is fated to be ultimately agreed to by all who investigate, is what we mean by the truth, and the object represented in this opinion is the real. That is the way I would explain reality.
>
> (Peirce, 1877: 155)

Peirce believed that one day, at some point so far into the future that no one can see it, all researchers would come to a 'final conclusion' that is *the* truth, and to which our present truths approximate. Dewey (1938) was more concerned with the immediate future, and coined the term 'warranted assertion', which he trades in for the notion of truth (and prefigures more recent approaches to validity as argument that we discuss in Unit A10). A warranted assertion is a claim that appears reasonable because it is usually confirmed by further practice and inquiry. Such 'convergence of enquiry' is necessary in the short term for practical purposes, but even for Dewey it is always possible that we will discover new methods or new practices that produce results which give us a better handle on the world.

Validity theory occupies an uncomfortable philosophical space in which the relationship between theory and evidence is sometimes unclear and messy, because theory is always evolving, and new evidence is continually collected. The fact that so many articles and books on testing and assessment use statistics cannot have escaped your notice, but the service to which this evidence is put is not always clear in a larger picture of developing theories of language acquisition and testing.

Positivistic validity theory (emphasizing as it did the verifiability of nomological networks) and later the falsifiability of nomological networks passed away because it was increasingly realized that theory and observation cannot be kept apart. We see through our beliefs, and our beliefs change because of observation. They are not watertight categories.

Task A1.4

➤ What is truth? From your experience as a teacher and/or tester is there anything that you consider an unquestionable truth? If you answer yes, what are your reasons? If you answer no, what are the consequences for how you teach and test?

A1.3 CUTTING THE VALIDITY CAKE

Since Cronbach and Meehl, the study of validity has become one of the central enterprises in psychological, educational and language testing. Perhaps the most significant figure in this work since the 1970s is Samuel Messick. In perhaps the most important article on validity, Messick (1989: 20) wrote:

> Traditional ways of cutting and combining evidence of validity, as we have seen, have led to three major categories of evidence: content-related, criterion-related, and construct-related. However, because content- and criterion-related evidence contribute to score meaning, they have come to be recognized as aspects of construct validity. In a sense, then, this leaves only one category, namely, construct-related evidence.

Messick set out to produce a 'unified validity framework', in which different types of evidence contribute in their own way to our understanding of construct validity. Messick fundamentally changed the way in which we understand validity. He described validity as:

> an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.
>
> (Messick, 1989: 13)

In this view, 'validity' is not a property of a test or assessment but the degree to which we are justified in making an inference to a construct from a test score (for example, whether '20' on a reading test indicates 'ability to read first-year business studies texts), and whether any decisions we might make on the basis of the score are justifiable (if a student scores below 20, we deny admission to the programme).

Table A1.1 presents this major step in our understanding of validity. In the left column is the 'justification' for testing, which can take the form of evidence or

Table A1.1 Facets of validity (Messick, 1989: 20)

	Test interpretation	*Test use*
Evidential basis	Construct validity	Construct validity + Relevance/ utility
Consequential basis	Value implications	Social consequences

consequences of testing. In the first row is the 'function or outcome' of testing, composed of interpretation or use. These two 'facets' give Messick's four-way progressive validity matrix.

The evidential basis for test interpretation is construct validity, and the evidence to support score meaning may be sought from any source. In this view, all evidence supports or weakens the intended score meaning, or the inferences that the test designers intended to make from the scores. The evidential basis of test use is also construct validity, but with specific reference to the context for which the test is designed or used. For example, we might wish to ask whether a test is appropriate for a particular group of learners in a specific context. The consequential basis of test interpretation is concerned with the theory and philosophy underlying the test, and what labels the test designer gives to the constructs. Labels send out messages about what is important or 'valued' in performance on the test, and this is part of the intended meaning of the score. The consequential basis of test use is the social consequences of actually using the test. When the test takers get their scores, how are the scores used by those who receive them? What kinds of decisions are made? And what impact do these decisions have on the lives of those who take the test?

Messick did not intend the categories of Table A1.1 to be watertight. Indeed, he explicitly stated that the boundaries were 'fuzzy', and suggested that it might be read as a 'progressive matrix' from top left to bottom right, with each category including everything that had gone before but with additions: from construct validity, looking at construct validity in specific contexts, then theory, and then the social consequences of the testing enterprise.

Task A1.5 ★

Think of a test that you are familiar with, perhaps one that you prepare students for.

- What construct(s) is the test designed to measure? Whom is the test designed for? Is it really relevant and useful for them? What are the parts of the test called? Are certain parts of language ability given preference or more highly valued, and does this impact on how you teach?

- What are the consequences for learners who fail, or get a low grade, on this test?

There are other ways of cutting the validity cake. For example, Cronbach (1988) includes categories such as the 'political perspective', which looks at the role played by stakeholders in the activity of testing. Stakeholders would include the test designers, teachers, students, score users, governments or any other individual or group that has an interest in how the scores are used and whether they are useful for a given context. Moss (1992) thinks that this is very similar to Messick's consequential basis for test use.

Messick's way of looking at validity has become the accepted paradigm in psychological, educational and language testing. This can be seen in the evolution of the *Standards for Educational and Psychological Testing*. In the Technical Recommendations (APA, 1954) the 'four types' of validity were described, and by 1966 these had become the 'three types' of content, criterion and construct validity. The 1974 edition kept the same categorization, but claimed that they were closely related. In 1985 the categories were abandoned and the unitary interpretation became explicit:

> Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the score. The inferences regarding specific uses of a test are validated, not the test itself.
>
> (AERA et al., 1985: 9)

The 1999 Guidelines go even further:

> The following sections outline various sources of evidence that might be used in evaluating a proposed interpretation of test scores for a particular purpose. These sources of evidence may illuminate different aspects of validity, but they do not represent distinct types of validity. Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose. Like the 1985 Standards, this edition refers to types of validity evidence, rather than distinct types of validity.
>
> (AERA et al., 1999: 11)

★ Task A1.6

Imagine that you work in a large language school and one of your tasks is to place one hundred new students into appropriate classes on their day of arrival.

A test exists for this purpose, but there is no evidence to support the validity of the scores for its purpose. From the list below, which pieces of information would be most useful for your evaluation of this test? Rank-order their importance and try to write down how the information would help you to evaluate validity:

- analysis of test content
- teacher assessments of students after placement
- relationship to end-of-course test
- analysis of task types
- spread of scores
- students' affective reactions to the test
- analysis of the syllabus at different class levels
- test scores for different students already at the school.

➤ Can you think of any other pieces of information that would be useful for your evaluation?

While Messick's approach is now dominant in validity theory, there have been further developments within the field of language testing that we need to consider.

A1.3.1 Test usefulness

Bachman and Palmer (1996: 18) have used the term 'usefulness' as a superordinate in place of construct validity, to include reliability, construct validity, authenticity, interactiveness and practicality. They have argued that overall usefulness should be maximized in terms of the combined contribution of the 'test qualities' that contribute to usefulness, and that the importance of each test quality changes according to context.

Reliability is the *consistency* of test scores across *facets of the test*. Authenticity is defined as the relationship between test task characteristics, and the characteristics of tasks in the real world. Interactiveness is the degree to which the individual test taker's characteristics (language ability, background knowledge and motivations) are engaged when taking a test. Practicality is concerned with test implementation rather than the meaning of test scores (see Unit 8A for a detailed discussion).

The notion of test 'usefulness' provides an alternative way of looking at validity, but it has not been extensively used in the language testing literature. This may be because downgrading construct validity to a component of 'usefulness' has not challenged mainstream thinking since Messick.

A1.3.2 The validity cline

In a series of important papers, Chapelle (1998, 1999a, 1999b) has considered how validity theory has changed in language testing since it was conceived as a property of a test (Lado, 1961: 321). In her work, Chapelle has characterized three current approaches to validity.

The first is traditional 'trait theory'. For our purposes, a 'trait' is no different from the notion of a 'construct', as used by Cronbach and Meehl. It is assumed that the construct to be tested is an attribute of the test taker. The test taker's knowledge and processes are assumed to be stable and real, and the test is designed to measure these. Score meaning is therefore established on the basis of correspondence between the score and the actuality of the construct in the test taker.

At the other end of the cline is what Chapelle terms the 'new behaviourism'. In a behaviourist approach the test score is mostly affected by context, such as physical setting, topic and participants. These are typically called 'facets' in the language testing literature. In 'real world' communication there is always a context – a place where the communication typically takes place, a subject, and people who talk. For example, these could be a restaurant, ordering food and the customer and waiter. According to this view, if we wish to make an inference about a learner's ability to order food, the 'real world' facets should be replicated in the test as closely as possible, or we are not able to infer meaning from the test score to the real world criterion.

This approach is typified in the work of Tarone (1998), in which it is argued that performance on test tasks varies (within individuals) by task and features or facets of the task. She argues that the idea of a 'stable competence' is untenable, and that 'variable capability' is the only defensible position. In other words, there are no constructs that really exist within individuals. Rather, our abilities are variable, and change from one situation to another.

Fulcher (1995) and Fulcher and Márquez Reiter (2003) have shown that in a behaviourist approach, each test would be a test of performance in the specific situation defined in the facets of the test situation. 'Validity' would be the degree to which it could be shown that there is a correspondence between the real-world facets and the test facets, and score meaning could only be generalized to corresponding real world tasks.

Trait theory and behaviourism are therefore very different in how they understand score meaning, and we can understand this in terms of the concept of 'generalizability'. Let us look at two extreme examples that will help make this clear.

Example 1: We design a test in which learners are presented with written text that contains a number of errors judged to be typical of learners of English as a second language. The test takers are asked to underline the errors, and write a correction. The score reflects the number of identified and corrected errors. From the score we make an inference about a learner's ability to write in English.

Example 2: We design a reading test in which learners are asked to read a car maintenance manual and complete process flow charts that show how to replace a clutch. The score reflects the level of success in completing the flow chart accurately. From the score we make an inference about a learner's ability to read a car maintenance manual to successfully replace a clutch.

In the first example we generalize the meaning of the score from a very specific error correction task to an ability to write – perhaps in any context and for any purpose! The claim being made is that error correction is a key part of the construct 'writing ability' and can predict success in the real world. Whether this could be supported with empirical evidence is a validity question. However, the underlying issue is important: in any test we can use only a small number of tasks or items, but we want to draw conclusions from the test scores that can generalize well beyond the sample of tasks or items in the test. Compare this with the second example. Here the score meaning is very limited. It has minimum generalizability, only to doing a very similar task in a non-test situation.

In practice we wish to be able to generalize score meaning from a limited number of tasks, but we acknowledge that the score from any particular test cannot be used for any purpose in the real world.

Task A1.7

Consider a test that you are familiar with, particularly one that many learners take, such as one of those produced by Educational Testing Service (ETS) (www.ets.org), Cambridge ESOL (http://www.cambridgeesol.org/index.htm), or some other testing agency.

➤ Who is the target population for the test?
What does the testing agency suggest the scores can be used for?
What task or item types are contained in the test?

➤ Do you think it reasonable to generalize from the scores to the suggested uses?

A more pragmatic stance is possible, however. Chapelle (1998: 34, 44) describes an *interactionist* understanding of score meaning as 'the result of traits, contextual features, and their interaction' and says that 'performance is viewed as a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts'. In this approach we acknowledge that the test contains only a sample of the situation or situations to which we wish to generalize. Part of investigating the validity of score meaning is therefore collecting evidence to show that the sample is domain-relevant, and predictive of the wider range of abilities or performances that we wish to say something about.

A1.3.3 Pragmatic validity

What we learn from the different approaches and definitions of validity is that validity theory itself is changing and evolving. We also learn that the things we look at to investigate validity may change over time. Similarly, our understanding of the validity of test use for a particular purpose is dependent upon evidence that supports that use, but the evidence and arguments surrounding them may be challenged, undermined or developed, over time.

What we call pragmatic validity is therefore dependent upon a view that in language testing there is no such thing as an 'absolute' answer to the validity question. The role of the language tester is to collect evidence to support test use and interpretation that a larger community – the stakeholders (students, testers, teachers and society) – accept. But this truth may change as new evidence comes to light. As James (1907: 88) put it, 'truth *happens* to an idea' through a process, and 'its validity is the process of its valid-*ation*' (Italics in the original).

The language tester cannot point to facts and claim a test valid. There are many possible interpretations of facts. What he or she has to do is create an argument that best explains the facts available. It is interesting to note that we talk of validity 'arguments' – a topic that we return to in Unit 10. The word 'argument' implies that there will be disagreement, and that there will be other interpretations of the facts that challenge the validity argument. 'Disagreements are not settled by the facts, but are the means by which the facts are settled' (Fish, 1995: 253). This is entirely in keeping with, but an expansion of, Messick's (1989) view that at the heart of validity was investigating alternative hypotheses to explain evidence collected as part of the validation process.

In a pragmatic theory of validity, how would we decide whether an argument was *adequate* to support an intended use of a test? Peirce (undated: 4–5) has suggested that the kinds of arguments we construct in language testing may be evaluated through *abduction*, or what he later called *retroduction*. He explains that retroduction is:

> the process in which the mind goes over all the facts of the case, absorbs them, digests them, sleeps over them, assimilates them, dreams of them, and finally is prompted to deliver them in a form, which, if it adds something to them, does so not only because the addition serves to render intelligible what without it, is unintelligible. I have hitherto called this kind of reasoning which issues in explanatory hypotheses and the like, *abduction*, because I see reason to think that this is what Aristotle intended to denote by the corresponding Greek term 'apagoge' in the 25th chapter of the 2nd Book of his Analytics. But since this, after all, is only conjectural, I have on reflexion decided to give this kind of reasoning the name of *retroduction* to imply that it turns back and leads from the consequent of an admitted consequence, to its antecedent. Observe, if you please, the difference of

> meaning between a *consequent*, the thing led to, and a *consequence*, the general fact by virtue of which a given antecedent leads to a certain *consequent*.

In short, we interpret facts to make them meaningful, working from the end to the explanation. In order to understand this more clearly, we will relate it to the stories of Sir Arthur Conan Doyle, for it is 'abduction' or 'retroduction' that is at the heart of every single Sherlock Holmes story ever written.

Task A1.8

★

Read this extract from *Silver Blaze* (Roden, 2000):

> 'We have here the explanation of why John Straker wished to take the horse out on to the moor. So spirited a creature would have certainly roused the soundest of sleepers when it felt the prick of the knife. It was absolutely necessary to do it in the open air.'
>
> 'I have been blind!' cried the colonel. 'Of course that was why he needed the candle and struck the match.'
>
> 'Undoubtedly. But in examining his belongings I was fortunate enough to discover not only the method of the crime but even its motives. As a man of the world, Colonel, you know that men do not carry other people's bills about in their pockets. We have most of us quite enough to do to settle our own. I at once concluded that Straker was leading a double life and keeping a second establishment. The nature of the bill showed that there was a lady in the case, and one who had expensive tastes. Liberal as you are with your servants, one can hardly expect that they can buy twenty-guinea walking dresses for their ladies. I questioned Mrs. Straker as to the dress without her knowing it, and, having satisfied myself that it had never reached her, I made a note of the milliner's address and felt that by calling there with Straker's photograph I could easily dispose of the mythical Derbyshire.
>
> 'From that time on all was plain. Straker had led out the horse to a hollow where his light would be invisible. Simpson in his flight had dropped his cravat, and Straker had picked it up – with some idea, perhaps, that he might use it in securing the horse's leg. Once in the hollow, he had got behind the horse and had struck a light; but the creature, frightened at the sudden glare, and with the strange instinct of animals feeling that some mischief was intended, had lashed out, and the steel shoe had struck Straker full on the forehead. He had already, in spite of the rain, taken off his overcoat in order to do his delicate task, and so, as he fell, his knife gashed his thigh. Do I make it clear?'
>
> 'Wonderful!' cried the colonel. 'Wonderful! You might have been there!'

> 'My final shot was, I confess, a very long one. It struck me that so astute a man as Straker would not undertake this delicate tendon-nicking without a little practice. What could he practise on? My eyes fell upon the sheep, and I asked a question which, rather to my surprise, showed that my surmise was correct.'

➤ What do you think are the key elements of Holmes's method? See if you can write down one or two principles that he uses to make facts meaningful.

The stories of Sherlock Holmes are gripping because the detective holds to a key principle: one eliminates alternative explanations, and the one that is left, however unlikely, is the most adequate. In language testing, the most adequate explanation is that which is most satisfying to the community of stakeholders, not because of taste or proclivity, but because the argument put forward has the same characteristics as a successful Sherlock Holmes case. And in language testing, the validity method is the same: it involves the successful elimination of alternative explanations of the facts.

In order to conduct this kind of validity investigation a number of criteria have been established by which we might decide which is the most satisfying explanation of the facts:

Simplicity, otherwise known as Ockham's Razor, which states: 'Pluralitas non est ponenda sine necessitate', translated as: 'Do not multiply entities unnecessarily.' In practice this means: the least complicated explanation of the facts is to be preferred, which means the argument that needs the fewest causal links, the fewest claims about things existing that we cannot investigate directly, and that does not require us to speculate well beyond the evidence available.

Coherence, or the principle that we prefer an argument that is more in keeping with what we already know.

Testability, so that the preferred argument would allow us to make predictions about future actions, behaviour, or relationships between variables, that we could investigate.

Comprehensiveness, which urges us to prefer the argument that takes account of the most facts and leaves as little unexplained as possible.

Task A1.9

Read more of the Sherlock Holmes story in the previous text box, available online at http://www.related-pages.com/sherlockholmes/showbook.asp?bookid=4&part=1&chapter=1.

Imagine that Holmes had concluded that the facts could only be interpreted through a theory that aliens had taken over Straker's body – in fact two aliens, who were fighting for control of his mind, which would account for the double life. And that his death had been caused by a third alien who had inhabited a horse in order to kill the other two aliens as a punishment for crimes on another world. In order to do this, the host body needed to be destroyed. The last alien, having accomplished his plan, left the horse and headed back into space.

➤ How would this theory violate the principles of simplicity, coherence, testability and comprehensiveness?

When you have completed this task, you will have discovered why the argument is not adequate, whereas in the story the argument is adequate, because it meets accepted *criteria for the evaluation of arguments.*

We conclude this section by reviewing the key elements of a pragmatic theory of validity:

1 An adequate argument to support the use of a test for a given purpose, and the interpretation of scores, is 'true' if it is acceptable to the community of language testers and stakeholders in open discussion, through a process of dialogue and disagreement.
2 Disagreement is an essential part of the process in investigating alternative hypotheses and arguments that would count against an adequate argument.
3 There are criteria for deciding which of many alternative arguments is likely to be the most adequate.
4 The most convincing arguments should start at the end point of considering the consequences of testing, and working backwards to test design.

Summary

Since the 1980s, validity inquiry has moved away from positivistic trait theory to include not only context but the utility of tests for the particular purpose for which they are designed. Nevertheless, there has been growing concern with the context of testing, and how test method is related to the target domain to which we wish to generalize. This is linked to the interest in the consequences of test use and the extent to which we should say what test scores should not be used for. Hence validity has been forced to consider the social and political reasons for test design and score use (Davidson and Lynch, 2002; Shohamy, 2001).

In this Unit we have looked at the development of validity theory from the early tripartite definition of content, criterion and construct validities to the present unitary interpretation of validity. We have defined key terms related to validity, and considered what we mean by constructs and construct validity. In this discussion we have seen that it is necessary to look at epistemological questions about the nature of truth, and ask the difficult question about whether our constructs are human creations, or ‘true’ in the sense that they have a separate existence in the real world.

Finally, we have outlined a pragmatic theory of validity that provides a backdrop to the treatment of other themes in this book.