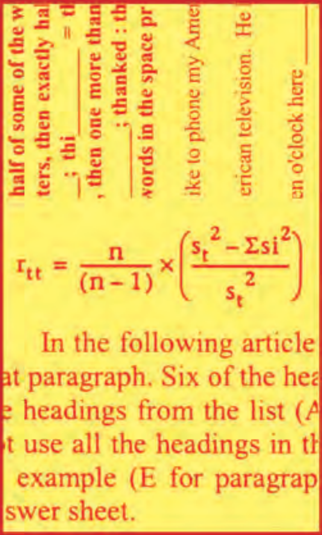
Research and Practice in Applied Linguistic
Series Editors: Christopher N. Candlin and David R. Hall

Language Testing and Validation

An Evidence-Based Approach



half of some of the w
ters, then exactly hal
_; thi ___ = th
, then one more than
___; thanked : th
words in the space pr
ike to phone my Ame
erican television. He
en o'clock here ___

$$r_{tt} = \frac{n}{(n-1)} \times \left(\frac{s_t^2 - \Sigma si^2}{s_t^2} \right)$$

In the following article
at paragraph. Six of the hea
e headings from the list (A
t use all the headings in th
example (E for paragrap
swer sheet.

Cyril J. Weir

Contents

General Editors' Preface

Research and Practice in Applied Linguistics is an international book series from Palgrave Macmillan which brings together leading researchers and teachers in Applied Linguistics to provide readers with the knowledge and tools they need to undertake their own practice-related research. Books in the series are designed for students and researchers in Applied Linguistics, TESOL, Language Education and related subject areas, and for language professionals keen to extend their research experience.

Every book in this innovative series is designed to be user-friendly, with clear illustrations and accessible style. The quotations and definitions of key concepts that punctuate the main text are intended to ensure that many, often competing, voices are heard. Each book presents a concise historical and conceptual overview of its chosen field, identifying many lines of enquiry and findings, but also gaps and disagreements. It provides readers with an overall framework for further examination of how research and practice inform each other, and how practitioners can develop their own problem-based research.

The focus throughout is on exploring the relationship between research and practice in Applied Linguistics. How far can research provide answers to the questions and issues that arise in practice? Can research questions that arise and are examined in very specific circumstances be informed by, and inform, the global body of research and practice? What different kinds of information can be obtained from different research methodologies? How should we make a selection between the options available, and how far are different methods compatible with each other? How can the results of research be turned into practical action?

The books in this series identify some of the key researchable areas in the field and provide workable examples of research projects, backed up by details of appropriate research tools and resources. Case studies and exemplars of research and practice are drawn on throughout the books. References to key institutions, individual research lists, journals and professional organizations provide starting points for gathering information and embarking on research. The books also include annotated lists of key works in the field for further study.

The overall objective of the series is to illustrate the message that in Applied Linguistics there can be no good professional practice that isn't based on good research, and there can be no good research that isn't informed by practice.

Christopher N. Candlin
Macquarie University, Sydney
and Open University, UK

David R. Hall
Macquarie University, Sydney

Abbreviations

AAAL	American Association of Applied Linguistics
ACTFL	American Council on the Teaching of Foreign Languages
AERA	American Educational Research Association
AERT	Advanced English Reading Test
AILA	International Association of Applied Linguistics
ALTE	Association of Language Testers in Europe
ANOVA	Analysis of Variance
APA	American Psychological Association
BAAL	British Association of Applied Linguistics
CALS	Centre for Applied Language Studies, University of Reading
CB	Computer-based
CBT	Computer-based Test
CEF	Common European Framework of Reference for Languages
CET	College English Test
CLA	Communicative Language Ability
CPE	Certificate of Proficiency in English
CR	Criterion-referenced
CRTEC	Centre for Research in Testing Evaluation and Curriculum
CUEFL	Communicative Use of English as a Foreign Language
EALTA	European Association for Language Testing and Assessment
EAP	English for Academic Purposes
EFL	English as a Foreign Language
ELBA	English Language Battery
ELT	English Language Teaching
EPTB	English Proficiency Test Battery
EPQ	Eysenck Personality Questionnaire
ERIC	Educational Resources Information Center
ESL	English as a Second Language
ESOL	English for Speakers of Other Languages
ESP	English for Specified Purposes
ETS	Educational Testing Service
FCE	First Certificate in English
FL	Foreign Language
FSI	Foreign Service Institute
GCE	General Certificate of Education
GCSE	General Certificate of Secondary Education

GEPT	General English Proficiency Test
IATEFL	International Association for Teaching English as a Foreign Language
IELTS	International English Language Testing System
ILR	Interagency Language Roundtable
ILTA	International Language Testing Association
IRT	Item Response Theory
JMB	Joint Matriculation Board (Northern Universities)
KR	Kudar Richardson
L1	First Language
L2	Second Language
LTM	Long-term Memory
LTRC	Language Testing Research Colloquium
LTTC	Language Training and Testing Centre
LTU	Language Testing Update
MCQ	Multiple-choice Question
MFR	Multi-faceted Rasch Analysis
MTMM	Multi-trait, Multi-method
NCME	National Council on Measurement in Education
NNS	Non-native Speaker
OPI	Oral Proficiency Interview
PBT	Paper-Based Test
PLAB	General Medical Council's Professional and Linguistic Assessments Board (Test of overseas doctors' language proficiency), United Kingdom
RSA	Royal Society of Arts
SAQ	Short-answer Question
SATD	Student Achievement Test Development
SD	Standard Deviation
SEM	Standard Error of Measurement
STM	Short-term Memory
SLA	Second Language Acquisition
SOPI	Simulated Oral Proficiency Interview
SPSS	Statistical Package for the Social Sciences
TEEP	Test of English for Educational Purposes
TEM	Test for English Majors
TL	Target Language
TLU	Target Language Use
TOEFL	Test of English as a Foreign Language
TOEIC	Test of English for International Communication
TSE	Test of Spoken English
TSWE	Test of Standard Written English
TWE	Test of Written English
UCLES	University of Cambridge Local Examinations Syndicate

UCRN	University of Cambridge Research Notes
UK	United Kingdom
USAID	United States Agency for International Development

Introduction

In language testing we are concerned with the extent to which a test can be shown to produce scores that are an accurate reflection of a candidate's ability in a particular area, e.g., careful reading to extract main ideas from a text, writing an argumentative essay, breadth of vocabulary knowledge, or spoken interaction with peers. It demands an understanding of both trait and method. Trait is concerned with the underlying constructs/abilities we wish to measure in students, the *what* of language testing. Method deals with the *how*, the instruments we develop to provide us with the information about these construct(s).

Test validation is the process of generating evidence to support the well-foundedness of inferences concerning trait from test scores, i.e., essentially, testing should be concerned with evidence-based validity. Test developers need to provide a clear argument for a test's validity in measuring a particular trait with credible evidence to support the plausibility of this interpretative argument (see Kane 1992). This is similar in a number of respects to a defence lawyer acting in the courtroom. As we will see below, this necessarily involves providing data relating to *context-based, theory-based* and *criterion-related* validities, together with the various *reliabilities*, or '*scoring validity*' as we prefer to call it.

Testing also has an ethical dimension in so far as it affects people's lives (see Davies (ed.) 1997). This leads us into the area of consequential validity where we are concerned with a test's impact on individuals, institutions and society, and with the use that is made of test results. Getting it right, ensuring test fairness, is a necessity not an ideal for testing. In developing assessment tools a decision must be taken on what is criterial in the particular domain under review, and this decision and the test measures used for operationalizing it must be ethically defensible. Test developers must be made accountable for their products.

Language testing is not just about creating the instruments for data generation – as it may seem from a number of practical books on the market, which deal principally with the mechanics of test production. Test development needs to go deeper than this, even when these are low-stake tests for use in the classroom for formative purposes. We want to show that testing must always be concerned with evidence-based validity, i.e., the relationships between the testing instrument and the construct(s) it attempts to measure.

The core of this book is concerned with exploring a framework for establishing the validity of the interpretation of scores on tests produced by Exam Boards or by teachers for use in their classrooms. This is what testing should be concerned with. Until it is, we can have little confidence in our interpretation of the test scores that are available to us. We offer below a blueprint of the types of evidence we must provide if we are to justify the correctness of our interpretations of abilities from test scores. Though specifically framed with English for Speakers of Other Languages (ESOL) in mind, the blueprint has implications for all forms of educational assessment.

This book follows the rationale and structure of the *Research and Practice in Applied Linguistics Series* in first providing a theoretical overview of the field, followed by detail of how this works in practice and then suggesting focuses and methods for researching key areas. In Part 1 we map out the types of validation evidence we need to provide if we are to have any confidence that the results of performance on a test give us an accurate picture of the underlying abilities or constructs we are attempting to measure. In Part 2 we unpack validity further in relation to actual examples and procedures taken from tests from around the world and provide an evidence-based validity framework for asking questions of any exam or form of assessment. In Part 3 we suggest a number of research activities, which will generate data on whether a test matches up to various criteria in the framework. Lastly, in Part 4, we detail a number of electronic and paper-based resources.

The first chapter sets the scene by tracing the development of language tests over the last century. It will attempt to describe the different stages in Western approaches to language testing, variously labelled by Bernard Spolsky (ed., 1978: v–x) as the 'pre-scientific' which lasted up to the Second World War and the 'psychometric-structuralist' that took us into the 1970s in Britain (but, as we will see, much later in the USA). Finally, we deal with the 'psycholinguistic-sociolinguistic' era covering the late 1970s until the present day in Britain (but really only taking off in the 1990s in the USA). More provocatively, these stages were described by Keith Morrow (1979) as 'the Garden of Eden, the Vale of Tears and the Promised Land'.

1
Language Testing Past and Present

Language tests from the distant past to the present are important historical documents. They can help inform us about attitudes to language, language testing and language teaching when little alternative evidence of what went on in the bygone language classroom remains. Seeing where we have come from also helps us better understand where we are today. The Cambridge ESOL Certificate of Proficiency in English (CPE) has by far the longest track record of any serious EFL examination still in existence, so it is a particularly useful vehicle for researching where we have come from in European approaches to language teaching and testing over the last century. We will trace some significant events in its history to exemplify the developments in the field during that period (see Weir 2003 for a full history of the CPE).

1.1 The Cambridge Proficiency Examination 1913–1945: 'The Garden of Eden', 'the pre-scientific era'

Weir (2003: 2) describes how Cambridge's formal entry into testing the English of foreign students took place in 1913, when it first offered the Certificate of Proficiency in English (CPE). The examination was based on the traditional, essay-based, native-speaker language syllabus including an English literature paper, the same as that sat by native speakers for university matriculation, and an essay, but also a compulsory phonetics paper, a grammar section and translation from and into French and German. These were complemented by an oral component with dictation, reading aloud and conversation.

The emphasis in this early pre-scientific era was thus on language use, though some attention was paid to form in the grammar and phonetics sections. The 'scientific' issue of test reliability was still relatively little understood, at least outside the United States (see Spolsky 1995) and the notion of the 'connoisseurship' of an elite group of examiners prevailed. All was thought to be well in this testing Garden of Eden.

1913 CPE Examination

(i) Written:	
(a) Translation from English into French or German	2 hours
(b) Translation from French or German into English, and questions on English Grammar	2 ½ hours
(c) English Essay	2 hours
(d) English Literature (The paper on English Language and	3 hours
Literature [Group A, Subject 1] in the Higher Local Examination)	1½ hours
(e) English Phonetics	
(ii) Oral:	
Dictation	½ hour
Reading and Conversation	½ hour

The 1913 test corresponded closely to the contents of Sweet's (1899) *The Practical Study of Languages: A Guide for Teachers and Learners* (see Howatt 1984 for details) and mirrored a concern with pronunciation as well as translation. Phonetics occupied a central position in the field of linguistics and language studies which was to survive until the 1960s in tests such as the English Language Battery Version A (ELBA) and the English Proficiency Test Battery (EPTB) used in university admissions (see Davies 2005 for a detailed account of these exams) and even later in the Professional and Linguistic Assessments Board (PLAB) test for overseas doctors wishing to practise in Britain. Grammar translation as a basis for testing proficiency was also to endure into the 1970s in most foreign language testing in the UK and still lingers on in the university sector. In contrast, the testing of English as a foreign language was to progress more quickly.

It is also interesting to note that an oral test (reading aloud and conversation) with associated dictation, was present in an international English as a Foreign Language (EFL) test at such an early stage. This multi-componential approach with a variety of discrete point, integrative and communicative tasks was to differentiate the Cambridge main suite examinations from most of its competitors through the twentieth century. It marks a British/European preoccupation with the trait, with *what* we are testing, as against an American preference for the method, the *how* of testing. This contrast was to last throughout the twentieth century until the Test of English as a Foreign Language (TOEFL) Next Generation programme.

Weir (2003: 14) points out how the approach in the first half of the century was to aim for construct validity and work on reliability, 'rather than through the single-minded pursuit of objectivity seriously curtail what CPE would be able to measure. A valid test that might not present perfect psychometric qualities was preferred to an objective test which though

Concept 1.1 Reliability and validity: competing paradigms in test development?

In these early days of language testing, reliability and validity were often seen as dichotomous concepts, a question of where priorities were to be placed. The cardinal guiding principle for Cambridge was construct validity, i.e., appropriateness in what was being measured, followed closely by utility for the teaching community. This does not mean they did not seek to achieve reliability, i.e., consistency of measurement, but reliability was not the overriding determinant of what went into the examination. According to Spolsky (1995), until the work of Roach in the 1940s on improving rater reliability, they appear to have remained relatively immune to psychometric influences from across the Atlantic.

always reliable might not measure that much of value, e.g., not test speaking or writing.'

In America the reverse was true and some aspects of validity were sometimes sacrificed in the pursuit of reliability. It is only with the recent development of TOEFL Next Generation that an attempt has been made to redress the situation by focusing on test activities more relevant to the demands of real-life academic study. Similarly in mainstream education in the USA, there is now increasing public concern over several aspects of validity of a number of the standardized tests that proliferate in school assessment despite their undoubted claims to reliability, i.e., measurement consistency.

We return to the issues of validity in Chapters 2, 3 and 4.

1.2 Developments in the 1960s: the move towards a language-based examination

Concept 1.2 Language tests should only test language

In the early 1960s we see the beginnings of a critical shift in the language testing tradition in Britain towards a view that language might be divorced from testing literary or cultural knowledge. It is thus possible in this period to date the start of a gradual but critical change of the English language examination to one which focuses on language as against an assortment of language, literature and culture.
(Weir 2003: 17–18)

Up to this point, the case for a language-based test had been hampered by the desire of linguists to gain academic respectability and recognition for language degree programmes in the older universities by injecting a heavy dose of literature and culture into their courses and examinations.

Weir (2003:19) describes how:

> candidates still have to take two other papers in addition to the compulsory 'English Language' paper. However, unlike the previous major revision in 1953, candidates can choose both 'Use of English' and 'Translation from and into English' as two additional papers, which means they do not have to take anything from (b) 'English Literature' or its alternatives.

1966

Oral: Dictation, Reading and Conversation
Written: Candidates must offer (a) English Language and **two** other papers chosen from (b), (c) or (d). No candidate may offer more than one of the alternatives in (b).

(a) English Language (composition and a passage or passages of English with language questions. The choice of subjects set for composition will include some for candidates who are specially interested in commerce)	(3 hours)
(b) Either English Literature Or Science Texts Or British Life and Institutions Or Survey of Industry and Commerce	(3 hours)
(c) Use of English	(3 hours)
(d) Translation from and into English	(3 hours)

In section (b) of the Use of English paper 3 option, multiple-choice items are introduced. This marks a growing interest in improving the reliability of the test overall, at least in terms of the internal consistency of the discrete item components (see Chapter 9). The more consistent the items were with each other in terms of how candidates performed on them, the higher this internal reliability. Spolsky (1978), in line with wider developments in the fields of statistics and linguistics, labelled this the 'psychometric-structuralist' era and Morrow (1979) 'The Vale of Tears'. The latter title was a reaction to an obsessive pursuit of objectivity, not just in tests of micro-linguistic knowledge (e.g., vocabulary) but also, for example, in the Multiple-Choice Question (MCQ) structure and written expression section in TOEFL. This indirect measure was used as an estimate of academic writing ability until the introduction of the bolt-on Test of Written English (TWE) paper in response to consumer wishes in 1986. Breaking language down into its elements also fitted well with the immediate constituent analysis of sentences in vogue with linguists in this period.

1.3 The 1975 and 1984 revisions: 'The Promised Land'?

The 1975 revisions saw the CPE examination taking a shape that, in its broad outline, is familiar to the Cambridge candidate of today and largely represents the content coverage of language tests at this level across the world. Weir (2003: 24) describes how

> the new CPE listening, reading and speaking tests in particular represented major developments on the 1966 revision and echoed the burgeoning interest in communicative language teaching in the 1970s; an increasing concern with language in use as against language as a system for study.... The 1970s saw a change from teaching language as a system to teaching it as a means of communication as set out and discussed in Widdowson (1978).

In the UK it was reflected in the teaching and publications emerging from CALS at the University of Reading under the influence of Ron White, Don Porter, Keith Morrow and Keith Johnson, and at Lancaster University under the influence of Chris Candlin, Michael Breen and colleagues.

The increased reliance on multiple-choice formats (in papers 2–4) acknowledged the attention international examinations felt they must pay to the demands of objectivity. The concern to improve marker reliability, particularly from the 1980s onwards, also aimed to improve the dependability of the scores in productive tests (papers 1 and 5).

The direct connection between the exam and British culture was completely broken and this potential source of test bias much reduced.

Content of the 1975 Certificate of Proficiency in English

PAPER 1 Composition	(3 hours)
PAPER 2 Reading Comprehension	(1¼ hours)
PAPER 3 Use of English	(3 hours)
PAPER 4 Listening Comprehension	(30 minutes)
PAPER 5 Interview	(approx. 12 minutes)

Weir (2003: 26) describes how

> the five papers have replaced the old division of Oral and Written and indicate some movement to recognizing further the need to address the notion that language proficiency is not unitary but partially divisible. It was to take a number of American applied linguists rather longer to discard their firmly held convictions that language proficiency was unitary and that therefore it mattered little what was tested as long as it was done reliably (see Oller 1979).

During the 1980s and 1990s there was, however, a degree of convergence of views on testing internationally, helped in no small part by the growing influence of the Language Testing Research Colloquium, which annually brought together researchers and scholars interested in language testing from around the world. The birth of the journal *Language Testing* as a result of a weekend meeting of a small group of British testers at Lancaster University in 1980 (see Alderson and Hughes (eds.) 1981) was to promote further the exchange of views across the Atlantic. The advent of the Language Testing list-serve, a web-based discussion forum in the 1990s, similarly promoted the exchange of views and an understanding of different traditions. The growing acceptance, or at least recognition, of international standards for language testing spawned by the drawing up of the American Educational Research Association *et al.* (1974, 1985, 1999) standards made an equally positive contribution. Full details of links to all of these can be found in Part 4.

Now that the channels of communication are open and earlier entrenched positions have softened, the future development of the field will depend on clarifying, codifying and disseminating a framework for test development, administration and analysis that all test developers can buy into. The rest of this book explores what might go into such a framework.

Further reading

Spolsky (1995) is an impressive, scholarly history of the development of ESOL examinations in the USA and Britain, if somewhat predisposed to the psychometric orientation of the former.

Weir and Milanovic (eds.) (2003) gives a full history of the development over a century of a major international ESOL examination the CPE looked at from a British perspective with its humanistic/sociolinguistic leanings.

palgrave
macmillan

Cyril Weir provides an innovative approach to language testin
is comprehensive and accessible to MA students in Applied Lin
TESOL and EFL, and to practising language teachers.

Teachers need a framework of questions to enable them to interpret scores on tests accurately, Students' motivation, progress and life chances may be greatly affected by their language examination results. This innovative evidence-based approach offers teachers a solid theoretical and practical base which will empower them critically to evaluate both their own tests devised for the classroom and those provided by Examination Boards.

Part 1 maps out the types of validation evidence needed to give confidence that the results of performance on a test provide an accurate picture of the underlying abilities or constructs that are being measured.

Part 2 examines real examples and procedures taken from tests around the world and provides an evidence-based validity framework for asking questions of any exam or form of assessment.

Part 3 suggests a number of research activities, large and small-scale, for generating data on whether a test matches up to the various criteria in the framework. This will be particularly useful to MA/professional teaching students undertaking research as part of their studies, or to practising teachers keen to put the framework into action.

Part 4 backs up the discussion of research and practice with information on key electronic and paper-based resources.

Cyril Weir holds the Powdrill Chair in English Language Acquisition at the University of Luton, and is Director of the Centre for Research in English Language Learning and Assessment. He has taught and provided consultancy in language testing, evaluation and curriculum renewal in over fifty countries worldwide. He has published widely in the fields of testing and evaluation. His books include *Communicative Language Testing*, *Understanding and Developing Language Tests*, and, as co-author, *Reading in a Second Language* and *Evaluation in ELT*. He is on the editorial board of *Language Assessment Quarterly*.

ISBN 1-4039-1189-4
90101
9 781403 911896

Printed in Great Britain

www.palgrave.com